

Multi-modal Speech Recognition Workshop 2002



10-12 June 2002

North Carolina A&T State University, Greensboro, USA

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

Sponsored by

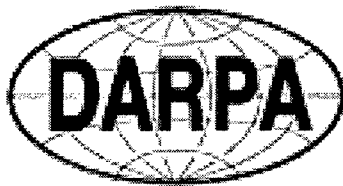
DARPA --- Defense Advanced Research Projects Agency
IMIS --- Center for Intelligent Mobile Information Systems
NC A&T State University

Editor : Sung H. Yoon

PROCEEDINGS OF
Multi-modal Speech Recognition Workshop

20 030506033

Multi-modal Speech Recognition Workshop 2002



10-12 June 2002
North Carolina A&T State University, Greensboro, USA

Sponsored by

DARPA --- Defense Advanced Research Projects Agency
IMIS --- Center for Intelligent Mobile Information Systems
NC A&T State University

Editor : Sung H. Yoon

Contents

v Workshop Committee

Session 1 Sensors and Sensor Fusion

- 1 **Physiological Sensors for Speech Recognition**
Mike Scanlon, Francis Fisher, Steve Chen, ARL
- 9 **A Surface Vibration Electromagnetic Speech Sensor**
Jonathan L. Geisheimer, Gene Grenaker, Georgia Tech Research
Institute; Scott Billington, Ittichote Chuckpaiwong, Georgia Institute of
Technology
- 13 **Evaluation of ASR Sensors**
Justin Taylor, Jason Heinrich, Jung H. Kim, Sung H. Yoon, North
Carolina A&T State University
- 17 **Multi-modal Sensory Fusion with Application to Audio-Visual
Speech Recognition**
Stephen M. Chu and Thomas S. Huang, University of Illinois at Urbana-
Champaign
- 21 **Alternative Speech Sensors for Military Applications**
Pete Fisher, U.S. Army Research Laboratory

Session 2 Audio-Visual Speech Recognition

- 27 **Development and Evaluation of Audio-Visual ASR: A Study on
Connected Digit Recognition**
Michael T Chan, Rockwell Scientific Company
- 32 **Multi-modal Dialog Systems Research at Illinois**
Stephen E. Levinson, Thomas S. Huang, Mark A. Hasegawa-Johnson,
Ken Chen, Stephen Chu, Ashutosh Garg, Zhinian Jing, Danfeng Li,
John Lin, Mohamed Omar, and Zhen Wen, University of Illinois
- 37 **Temporal Asynchronicity Modeling by Mproduct HMMs for Audio-
Visual Speech Recognition**
Satoshi Nakamura, ATR Spoken Language Translation Research
Laboratories
- 41 **Visual Speech Feature Extraction From Natural Speech for Multi-
modal ASR**
Sabri Gurbuz and John N. Gowdy, Clemson University
- 49 **Performance Analysis of Automatic Lip Reading Based on Inter-
Frame Filtering**
Jimyong Kim, Seongmo Park, Chonnam National University; Seungho
Choi, Dongshin University
- 53 **Robust Head Tracking Based on Hybrid Color Histogram and
Random Walk Kalman Filter**
Gwang-Myung Kim, Dongcheng Lin, Jung H. Kim, Sung H. Yoon,
North Carolina A&T State University
- 59 **The Validation of Military Callsign Intelligibility**
Celestine A. Ntuen & Misty Blue, North Carolina A&T State University
- 62 **Large Vocabulary Audio-Visual Speech Recognition**
C. Neti and G. Potamianos, IBM T.J. Watson Research Center

- 67 **Who? What? Where? How? Perceptually Aware User Interfaces**
Alex Waibel, Carnegie Mellon University
- 76 **Joint Audio-Visual Speech Recognition and CMU Audio-Visual
Speech Data Set**
Tsuhan Chen, Carnegie Mellon University
- Statistical Modeling of Data-fusion for Classifier Systems**
Jeff Blimes, University of Washington

Workshop Committee

Workshop Chairs

Dr. John Kelly, North Carolina A&T University

Dr. James D. Bass, DARPA/ITO

Program Committee

Dr. Jung Hyoun Kim, North Carolina A&T University

Dr. Michael Chan, Rockwell Scientific

Dr. Sung H. Yoon, North Carolina A&T University

Dr. Jung Hee Kim, North Carolina A&T University

Session Chairs

Dr. Pete Fisher, Army Research Lab

Dr. Michael Chan, Rockwell Scientific

Session 1:

Sensors and Sensor Fusion

Physiological Sensors for Speech Recognition

Mike Scanlon, Francis Fisher, Steve Chen

Abstract. *Systems designers are expressing greater interest in speech-based user interfaces for a variety of civilian and military applications. Such interfaces provide hands-free operation and a more natural way for humans to interact with systems. One difficulty with speech-based user interfaces is poor operation in noisy environments such as military operations. The Physiological Sensor, developed at ARL, is an example of an alternative sensor for automatic speech recognition. This sensor detects speech by measuring acoustic signals through the speaker's skin. While the signal produced is not typical of that from an airborne acoustic microphone, the possibility exists for using this sensor as a microphone. We investigate several possible methods for using the Physiological Sensor as a microphone for automatic speech recognition.*

1. Introduction

With recent advances in automatic speech recognition (ASR) technology has come an increased interest in applying this technology to the design of user interfaces. For a system being operated in a benign environment such as an interface can be based on commercial or custom software and an airborne acoustic microphone. However, most systems of this type are difficult or impossible to use in noisy environments such as those presented in military or industrial scenarios. In such cases we must find alternative ASR software or speech sensors in order to enhance operation in these environments. Efforts to improve operation in noisy environments by removing the noise from the microphone output have proven difficult without knowledge of the external noise source.

The Physiological Sensor, a medical sensor developed at Army Research Laboratory, is a device that physically couples to a patient to record medical information such as respiration and heartbeat. With some slight modifications to the electronics, ARL has converted this sensor to a microphone to be worn around the throat.

2. Physiological Sensor - Background

ARL has developed a new method to measure human physiology and monitor health and performance parameters. This consists of an acoustic sensor positioned inside a fluid-filled bladder in contact with the human body. Packaging the sensor in this manner minimizes

outside environmental interferences, and signals within the body are transmitted to the sensor bladder with minimal losses. This fluid-coupling technology comfortably conforms to the human body, and enhances the signal-to-noise-ratio (SNR) of human physiology to that of ambient noise. An acoustic sensor system can detect changes in a person's physiological status resulting from exertion or injuries such as trauma, penetrating wound, hypothermia, dehydration, heat stress, and many other conditions (or illnesses). Furthermore, a sensor contacting the torso, head, or throat region picks up the wearer's voice very well through the flesh, with fidelity sufficient to be used as an auxiliary microphone for communications or hands-free voice activation mechanism. Automatic speech recognition software, in conjunction with this enhanced body-coupling sensor, could improve mission performance by reducing false voice commands through improved SNR in noisy environments. Civilian technology transfer applications include clinical surveillance, medical transport, hospitals, and telemedicine applications. Fire, rescue, and police personnel may benefit from hands free voice communications with embedded health and performance monitoring [Scanlon, patents].

2.1 Sensor Description

The neck-band sensors shown in figures 1 and 2 consists of a housing, gel-coupling sack with sensor embedded within, neck strap, preamplifier, and battery pack with hardwired signal egress and push to talk button.. The headband sensor in figure 3 does not use a liquid coupling, but rather an acoustically conductive silicone rubber.

Data were collected at the side of the neck using a sensor of similar geometry to the sensor in figure 1 [Scanlon, 1998]. The test included a spoken word count from 1 to 10, then mouth breathing for the remainder of the data set. Naturally, the heartbeat is always present. The time and frequency representations are shown in figure 4. Figure 5 compares data from a B&K microphone in front of the speaker's mouth to that of a fluid-coupled physiological sensor held in contact with the neck by a strap. Data from both locations were taken simultaneously in a typical office environment. Comparing the amplitudes of the voice to the non-vocal ambient noise surrounding the voice gives approximately 40 dB SNR for the B&K airborne microphone, and approximately 75 dB SNR for the fluid-coupled sensor. The fluid coupling represents an

improvement of better than 30 dB in speech SNR with minimal waveform degradation, as observed by the similarity of spectrograms and by listening to the data through headphones.

Time (s)
Time (s)

The ability of body-coupled sensors to detect physiology and reduce background noise was investigated. A physiological sensor was attached to one side of a speaker's neck, and an omnidirectional electret microphone was placed in front of the mouth. Figures 6 and 7 show simultaneously collected breath and voice data before, during, and after a speaking subject is immersed in a C-weighted noise field of 105 dB (referenced to 20 micropascals) noise field. The person wearing the sensors repeatedly vocalized a 1 to 10 count between the times of 14- and 19-s, 25- to 33-s, 65- to 71-s, and 71- to 77-s, and vocalized "105 dB" between 47- and 50-s.

The boom microphone in figure 6 does not detect any voice during the high amplitude noise between 20- and 71-s. However, in figure 7, the counting is clearly visible throughout the loud noise with the body-coupled gel sensor. Playing the data collected through headsets, the listener could clearly hear and understand the spoken words from the gel sensor in 105 dB noise, but could not discern the presence of any speech in the boom microphone data.

3. Automatic Speech Recognition Using the Physiological Sensor

Army Research Laboratory (ARL) and Rockwell Sciences Center (RSC) have developed several experimental systems that use the Physiological Sensor as input to automatic speech recognition (ASR) systems. These efforts are discussed below.

3.1 RSC Integration & application of the Physiological Sensor

3.1.1 General Signal Characteristics of the Physiological Sensor

By coupling directly to the user's neck, the physiological sensor was able to achieve extraordinary signal to noise performance as compared to airborne acoustic microphone technologies. While providing significant rejection of ambient noise, the sensor was not entirely immune to ambient sound. For instance, it was quite possible to detect other persons speaking to the wearer of

the physiological sensor, though at greatly attenuated levels. Due largely to the method of transduction, the output signal of the ARL physiological sensor was significantly different from typical acoustic microphone signals. Specifically, higher frequencies tended to be significantly attenuated. Human listeners listening to the output signal of the physiological sensor indicated that the distortion was analogous to listening to a person in another room through a wall.

3.1.2 Physiological Sensor and Speech Recognizers

Because of the inherent distortions of speech associated with the ARL physiological sensor, many commercial, off-the-shelf ASR technologies, like IBM's ViaVoice, were unable to successfully recognize speech using the physiological sensor signals. Such recognizers often rely on Hidden-Markov Models of speech, where the models are pre-estimated using statistical methods and large databases of human speech. Such databases would have been collected with conventional airborne acoustic microphones, so any speaker-independent speech recognizer would have an inherent expectation about the signal characteristics of speech as normally acquired through airborne acoustic microphones. Hence, in performing speech recognition with the physiological sensor, speaker-dependent recognizers tended to work more reliably. As recommended by ARL, the initial speech recognition engine utilized was the Clamor engine, a dynamic-time-warping speech recognizer developed by the Lexicus business unit of Motorola. Clamor recorded templates of each word or phrase ("token") to be recognized as provided by the user (2 instances of each token were kept as matching templates). Performance with the Clamor recognition engine was adequate for discrete, speaker dependent recognition of up to several distinct tokens.

Later, Rockwell Science Center developed a speaker-dependent, Hidden-Markov Model based discrete speech recognizer for use with the physiological sensor. The HMM-based recognizer was designed using HTK, a product of the former Entropic Research Laboratories. Like the DTW-based Clamor recognizer, RSC's HMM-based recognizer provided discrete recognition for up to several distinct tokens. The key difference was that with an HMM-based recognizer, additional training samples could be used to re-estimate the speech models, and presumably build a more robust, statistically accurate model of each token as more and more training utterances were collected from the user. The refined HMM models should perform better, while still maintaining the same level of computational complexity. With the DTW approach, the use of additional user utterances for

recognizer training would necessarily increase the computational burden of speech recognition at runtime – the more templates that were collected, the longer each match would take.

In order to support rapid integration and testing of user interfaces involving the physiological sensor, it was integrated with Rockwell's Automatic Speech Recognition (ASR) Server technology. The ASR Server provided abstraction of an encapsulated speech recognition engine (Clamor was used for the physiological sensor) through a platform-neutral TCP/IP socket interface. Applications could be quickly designed to exploit speech recognition services of the ASR Server through a simplified protocol. The ASR Server could, in turn, provide speech recognition through either the physiological sensor, or a conventional acoustic microphone. The physiological sensor was demonstrated in conjunction with Rockwell's Multimodal Integrated Displays Testbed in early 1999.

In early 2000, RSC's HMM-based recognizer for the physiological sensor was integrated with RSC's Bimodal ASR Server. The Bimodal ASR Server employed a subset of the same client/server interface protocol used by the ASR Server; whereas the ASR Server encapsulated COTS acoustic speech recognizers, the Bimodal ASR Server encapsulated more experimental recognition technologies, including both the HMM-based recognizer for the physiological sensor, as well as the visual lip-tracking based speech recognizer described in elsewhere in this text. The physiological sensor and Bimodal ASR Server were demonstrated as components of Rockwell's Integrated Displays Testbed v2 in early 2000 [Vassiliou, 00]. As part of the demonstration, a user could dynamically switch between speech recognition using either the lip tracker or the physiological sensor.

The natural extension of this work would be development of a hybrid speech recognition technology that concurrently uses both the physiological sensor and the visual speech recognizer. The two technologies are uniquely complementary because while the visual speech recognizer leverages key visible features of speech articulator motion (vital for recognition of consonant sounds), it is unable to distinguish voiced from unvoiced speech, and indeed is fairly unsuitable for discrimination of vowel sounds from one another. On the other hand, because of its nearly direct coupling to the vocal tract, the physiological sensor is advantageously placed for detecting voicing and discriminating vowel sounds, while its ability to capture subtle acoustic transients of consonant production may be compromised by its body-coupled nature. The visual speech recognizer is already HMM based, so significant research opportunities exist for the development of appropriate

feature vectors and HMM topologies to integrate the two distinct signal streams (visual & acoustic).

3.1.3 Ergonomics

The physiological sensor was found to be generally comfortable to wear, though there were some issues with the design. One obvious problem was that users wearing a collared dress shirt could have problems fitting the physiological sensor band either above or under the collar. Generally, with a shirt collar closed, fitting the physiological sensor inside the collar band was not practical. Wearing the physiological sensor higher on the neck than a closed shirt collar tended to limit head movement. Possibly, a narrower band and smaller sensor capsule could help with these issues.

The neck band itself was fairly easy to secure due to the use of Velcro surfaces. The fabric of the neck band was of a dense weave, which could lead to the accumulation of perspiration under the neck band under some conditions. A thinner, more loosely woven fabric, perhaps an elastic one, might be helpful.

The physiological sensor was also compared to a similar COTS throat worn microphone product, the LASH II microphone distributed by Television Equipment Associates. While the LASH II did use a thinner, narrower, elastic collar band, the plastic hook assembly for closing and securing the LASH II was not as easy to use as the Velcro design of the ARL physiological sensor. Further, the LASH II design caused two rigid plastic nodes to be pressed against the user's throat, which could cause significant discomfort when worn over extended periods. In contrast, wearers generally did not find the ARL physiological sensor to increase in discomfort over time.

Some hesitance and psychological resistance to wearing the physiological sensor was also reported of prospective users. An obvious safety concern for any neck worn apparatus is the possibility of choking, either by accident or by assailants. Also, while head worn microphones of some styles have come to be socially acceptable to wearers and even fashionable or "cool" in certain contexts, the visual appearance of the neck worn physiological sensor was less acceptable to some users.

3.1.4 Physiological Sensor Integration Issues

In early 1999, Rockwell received first samples of the ARL Physiological Sensor technology. Early samples used a fairly large (~5"x3"x2") preamplification module, which was rather bulky and not well suited to bodyworn applications. Despite having a full metal casing, the

combination of physiological sensor and preamplification module was also susceptible to grounding problems, which would cause a strong 60Hz hum to be present in the output signal. The grounding problems were corrected in the next received prototype early in 1999 and the physiological sensor was successfully mated to a PC-based sound card using the line level input. Some speech recognizers are designed with the assumption that the microphone input of a sound card will be used for speech acquisition, so the user of line level input could have been an integration issue for some speech recognition technologies.

Newer versions of the physiological sensor supplied by ARL in late 1999 and early 2000 used a much smaller and lighter preamplification module (~2"x1"x.5") in a plastic rather than a metal housing. The new preamplification module was light enough to be carried with the user, and the signal level was suitable for use with the microphone inputs of typical PC sound cards. It also included a momentary push-to-talk switch. Conceptually, a push-to-talk switch is helpful in speech recognition applications because if the press and release events for the switch can be detected by the speech recognizer, then delimitation of user utterances becomes fairly easy. Also, the use of a push-to-talk switch helps to prevent false recognition (insertion) errors where extraneous noises or speech not intended for the recognizer are acquired by the transducer. In the case of the current versions of the physiological sensor though, the implementation of the push-to-talk switch is suboptimal for speech recognition. First, the switch is electromechanical and entirely embedded in the preamp module of the physiological sensor, so there is not a deterministic way (e.g. additional connector pin) for an attached device or computer to ascertain when the switch is pressed and released. The switch also induces significant transients in the sensor's output signal when it is pressed and released. Such transients in the speech signal are apt to confuse most existing speech recognition technologies. The workaround solution employed to address these issues was to keep the push-to-talk switch depressed at all times while using a speech recognition system, and to rely on other, external push-to-talk switch mechanisms that were more readily tracked by the Rockwell ASR Server. Additionally, because the push-to-talk switch was of a momentary-on design, additional external fixtures were required to keep the switch depressed.

For some applications, it was desirable for the user of the physiological sensor to be free to move about untethered. Attempts were made to connect the physiological sensor to a wireless microphone transmitter module (Audio Technica ATW-T75), but the output signal levels and impedance were found to be not fully compatible with the input stages available on the wireless transmitter.

Although a signal could be sent wirelessly, additional distortions were introduced, which ultimately degraded speech recognition accuracy.

RSC has provided ARL with recommendations for improvements to the design of future Physiological Sensor based microphones.

3.2 Army Research Laboratory

ARL has conducted two experiments using the Physiological Sensor as an input device for ASR. The first effort used the Entropic HTK as the automatic speech recognition (ASR) engine and compared the capabilities of the Physiological Sensor with an acoustic microphone. The second effort utilized Dragon Systems Naturally Speaking, a commercial ASR product to evaluate the possibility of using the Physiological Sensor with commercial speech engines.

All applications of the Physiological Sensor as a speech input device must take into account the difference in frequency response of this sensor as compared to a typical airborne acoustic microphone. This difference in frequency response typically precludes the use of acoustic language models provided with most ASR systems.

3.2.1 Physiological Sensor with Entropic HTK

For the experiment using HTK, ARL teamed with the United States Military Academy (USMA) to develop speech models appropriate for use with the Physiological Sensor [Bass, 99]. The Entropic HTK, a Hidden Markov Model based system, was chosen because it provides the flexibility required to adapt the internal configuration of the ASR engine for use with the Physiological Sensor.

The test consisted of trying to recognize one of 50 phrases using both an airborne sensor (microphone) and the Physiological Sensor. Two recognizers were used, each trained on one of the sensors being tested. Phrases consisted of two to ten words each, with a total of 153 unique words. Each test subject spoke the phrases in an environment that yielded speech to noise ratios of 0-, 3-, and 10dB SNR through the airborne sensor, while wearing both the airborne and physiological sensors.

Speech training and testing was conducted by USMA at their facilities. Training was performed using data collected from 21 subjects speaking the 50 phrases in a quiet environment. The result of the training is a speaker independent model for recognition of the 50 test phrases. Testing was then performed on data collected using 14 new subjects to speaking the 50 phrases in each of the given noise environments.

The results of this experiment are shown in tables 1 and 2. In all cases the Physiological Sensor and related recognizer outperformed the airborne acoustic sensor and related recognizer for the given noise levels. Further, the % accuracy of the Physiological Sensor degrades at a much lower rate with increased noise as compared to the airborne acoustic sensor.

3.2.2 Physiological Sensor with Dragon Naturally Speaking

In order to evaluate other possible application areas for the Physiological Sensor we decided to perform a limited test with a commercial ASR product. We selected Dragon Naturally Speaking for the test because we had considerable experience using this product. To simplify the experiment we used the same set of phrases as used with the HTK testing. One user trained the system using the standard user training session. In addition, all of the words in the command phrases were trained separately.

With this very limited data set, 50 phrases and one user, recognition rates were found to vary between about 60% and 80%. While not outstanding, this is a fairly good result considering that the ASR engine was developed for an airborne acoustic microphone. It should be noted that the worst recognition rates were obtained when the user removed and reattached the Physiological Sensor. We assume that changes in the sensor pressure and position are the cause for these variations. No tests were performed in the presence of noise.

3.2.3 Future Research and Experimentation

Experiments with the Physiological Sensor have demonstrated its capability to be used as a speech sensor for specially trained and configured ASR systems. The requirement for special configurations prevents the application of this sensor with many of the commercial ASR products on the market. Since the private sector is investing heavily in the development of these continually improving commercial ASR products it makes sense to leverage this effort. As a result, ARL will work to develop methods to convert the output of the Physiological Sensor into a signal that more closely approximates that of an acoustic sensor. If we can accomplish this then the Physiological Sensor should be suitable for use with any commercial ASR product. The resulting system would have the improved capabilities of the commercial ASR products with the noise rejection capability of the Physiological Sensor.

4. Summary, Conclusions (Lessons Learned), and Recommendations

Several areas exist to improve the operation of the Physiological Sensor as a microphone. The sensor already has good airborne noise rejection, but more can be done to limit the amount of airborne noise that couples to the sensor. An acoustic insulation material can be incorporated around the shroud of the sensor that contacts the skin to prevent the airborne noise from contacting the sensor's gel pad. Additionally, sensors could be mounted on both sides of the throat and their outputs summed simultaneously so that the speech would add constructively, whereas the noise would be reduced by common mode rejection. Since the vocal folds are not always symmetrical, the combined left and right signal may improve intelligibility through construction of an enhanced signal.

One potential problem in the application of the Physiological Sensor as an input to ASR systems is the substantial variation in signal due to changes in sensor pressure and position. We will research this issue in the future and attempt to minimize these effects in order to improve operation with ASR software.

Circuit modifications can be made to eliminate noises from switch activation, match impedance for interaction with other devices, and adjust the filtering to create a more accurate representation of the speech. The preamplifier used in all of the experiments described herein had a flat response, and did not enhance or boost the high frequencies that are lower in amplitude than the very dominant lower formants. Developing a non-linear amplifier (filter) can reduce the "through the wall" perception developed by some listeners, and may produce waveforms that better match what the commercial ASR engines expect. In addition, refinement of ergonomics and packaging would be worthwhile for maturing this technology into a product.

The physiological sensor has demonstrated exceptional capabilities for the detection of voice in high noise environments. In addition, the physiological parameters detected by this sensor provide health and performance indication, but might ultimately provide invaluable emotional or physiological data that can be used to adapt and optimize ASR algorithms under diverse situations. This is important in almost every military and civilian application. Acoustics can provide invaluable clues to help understand the interrelations between the soldier's physiology, the task at hand, the spoken word's intent, and the surrounding environment.

Areas requiring future research include the development of a user independent HMM model set to assist developers working with the Physiological Sensor, development of algorithms or filters to enhance operation of the sensor for use with commercial ASR products, and refinements in overall operation.

References

- [Bass, 99] J. Bass, M. Scanlon, T. Mills, and J. Morgan, "Getting Two Birds with One Phone: An acoustic sensor for both speech recognition and medical monitoring", *Acoustical Society of America*, November 1999.
- [Scanlon, 98] M. Scanlon, "Acoustic Sensor for Health Status Monitoring", *Proceedings of the 1998 Meeting of the IRIS Specialty Group on Acoustic and Seismic Sensing*, Volume II, pp. 205-222.
- [Scanlon, patents] M. Scanlon, "Sudden infant death syndrome (SIDS) monitor and stimulator", May 1996, U.S. Patent 5,515,865; "Motion and sound monitor and stimulator", Nov. 4, 1997, U.S. Patent 5,684,460; "Acoustic monitoring sensor", December 29, 1998, U.S. Patent 5,853,005.
- [Vassiliou, 00] M. Vassiliou, V. Sundareswaran, S. Chen, R. Behringer, C. Tam, J. McGee, "Integrated multi-modal human-computer interface and augmented reality for interactive display applications", *2000 SPIE Aerosense*, April 24-28, 2000, Orlando FL.



Figure 1: Gel sensor pad.

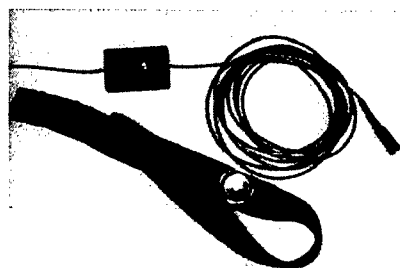


Figure 2: Neck assembly for voice.



Figure 3: Sensor in helmet headband.

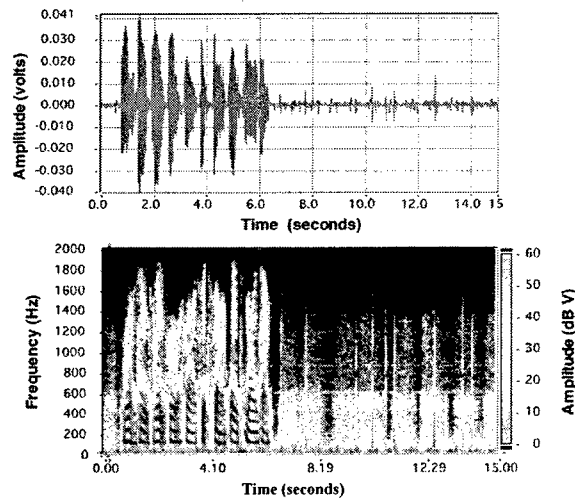


Figure 4: Fluid sensor held at throat for 1 to 10 voice count and mouth breaths.

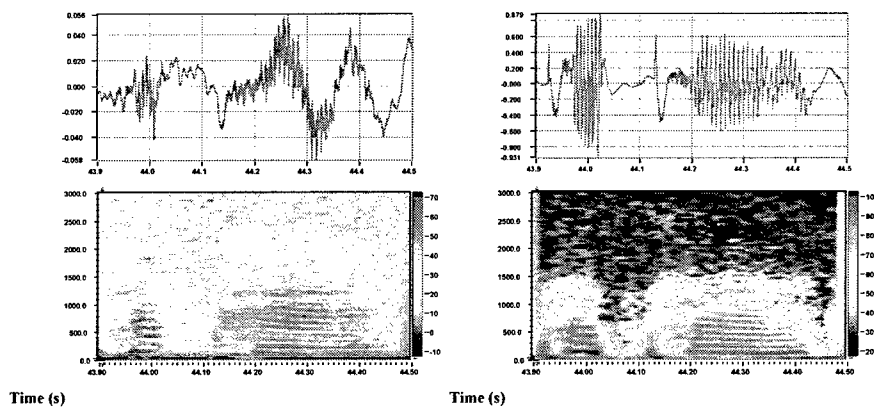


Figure 5: Comparison of spoken word "papa" taken with ambient microphone (left) and throat pad (right).

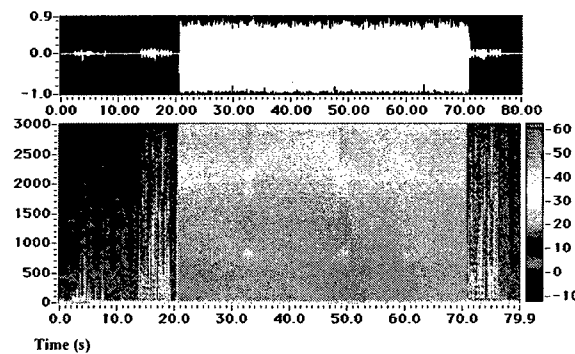


Figure 6: Boom microphone detecting voice.

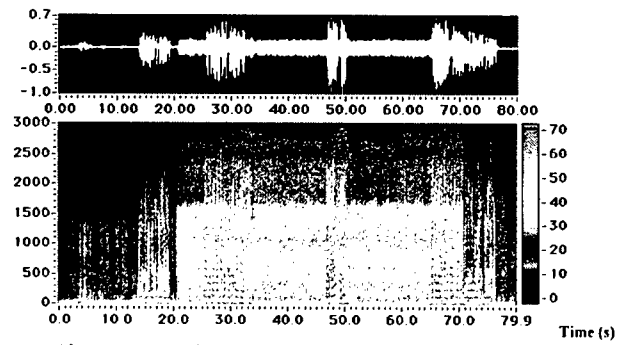


Figure 7: Gel sensor on neck detecting voice.

Table 1. Sentence Loop Language Model

Sentence Loop Model (% Perfect Sentence Recognition)		
dB	Airborne	Physiological
0	40.6	96.5
3	60.7	98.7
10	98.7	99.4

Table 2. Word Loop Language Model

Word Loop Model (% Perfect Sentence Recognition)		
dB	Airborne	Physiological
0	-0.1	39.6
3	12.8	50.5
10	51.5	66.8

A Surface Vibration Electromagnetic Speech Sensor

Jonathan L. Geisheimer, Eugene F. Grenaker, Scott A. Billington, Ittichote Chuckpaiwong

Abstract—As researchers continue to improve speech in noisy environments, more interest is being placed on sensors with modalities that can be fused with traditional acoustic sensors. The standard literature has shown that electromagnetic sensors can be used to detect glottal motion. Also, accelerometers placed on the throat and nasal areas have been used to detect skin surface vibrations corresponding to speech and that data has been used for noise reduction. The Georgia Tech Research Institute (GTRI) is transitioning a 24 GHz radar technology originally used for non-contact vital signs monitoring to a technology able to measure surface motion on the order of microns, which can detect skin surface vibrations corresponding to speech. The radar has been shown to measure the same motion as accelerometers using electromagnetic waves. This paper describes the theory and preliminary work in developing a surface vibration electromagnetic speech sensor to be used for noise reduction in conjunction with acoustic sensors.

Index Terms—radar, speech, noisy environments, sensor fusion.

I. INTRODUCTION

Every time a person speaks, the acoustical pressure waves from speech couple through many parts of the body, which causes structures such as the head, neck, chest, and face to vibrate. If a hand is placed on the chest or throat when speaking, these vibrations can be readily felt. The acoustic pressure waves due to speech have been translated to mechanical vibrations. This has been confirmed by various researchers who have looked at the head and chest vibrations in signers.¹ Other researchers have detected mechanical vibrations off of the neck using contact accelerometers and have been successful in using the resultant vibration signal to cancel noise when fused with acoustic data.^{2,3}

An electromagnetic-based sensor called the Glottal Electromagnetic Micropower Sensor (GEMS), developed at Lawrence Livermore National Laboratories,⁴ has been used to detect internal body vibrations. This sensor uses a low power, wideband pulsed radar that is able to penetrate through the body and detect glottal movement.⁵ It operates at microwave frequencies less than 3.0 GHz. In general, lower

microwave frequencies will achieve better penetration into the body.

The surface vibration electromagnetic speech sensor concept uses electromagnetic waves in the millimeter wave region to measure the slight vibrations of the body on the skin corresponding to human speech, down to micron levels of motion. At the proposed operational frequency of 35.0 GHz, the electromagnetic waves pass through clothes but do not penetrate into the body as does the GEMS sensor. The radar is detecting only surface vibrations and therefore directly measures the surface skin vibration and not the internal body structures. Since the device is directly picking up speech vibrations, it will be referred to as a “radar microphone”. A diagram of the concept is shown in Figure 1.

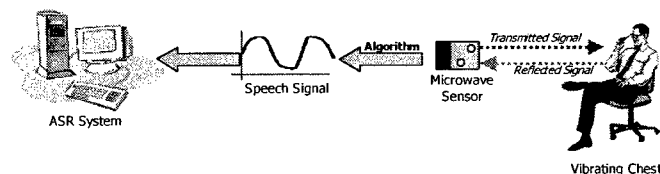


Figure 1. Radar Microphone concept

Referring to Figure 1, the radar microphone transmits a continuous wave (CW) electromagnetic signal towards the person's chest or neck area. Next, the signal is reflected back to the sensor where it is demodulated and converted to a baseband signal, sampled by an analog-to-digital converter, and then run through digital signal processing algorithms to convert the radar signal into displacement that correlates with the surface vibrations due to speech. The resultant speech signal can then be fused with other more traditional speech sensors and then passed on to an automatic speech recognition system if desired.

II. TECHNOLOGY BACKGROUND

The Georgia Tech Research Institute (GTRI) has been sensing small-scale biological motion using radar for almost 20 years, beginning with the Radar Vital Signs Monitor (RVSM). RVSM technology is able to detect both respiration and heartbeat signatures from individuals without contact. The first GTRI RVSM system was developed in the mid-1980s under sponsorship of the United States Department of Defense (DOD); a patent on the system was issued in 1992.⁶ This frequency modulated (FM) radar was used as a battlefield vital signs monitor. The system was tested on soldiers wearing a chemical or biological warfare suit to allow vital signs to be monitored without opening the suit and risking contamination of the subject.⁷

J. L. Geisheimer is with the Sensors & Electromagnetic Applications Laboratory at the Georgia Tech Research Institute, Atlanta, GA 30332 USA (telephone: 770-528-7690, e-mail: jon.geish@gtri.gatech.edu).

E. F. Grenaker is with the Sensors & Electromagnetic Applications Laboratory at the Georgia Tech Research Institute, Atlanta, GA 30332 USA (telephone: 770-528-7744, e-mail: gene.grenaker@gtri.gatech.edu).

S. A. Billington is with the Manufacturing Research Center at Georgia Tech, Atlanta, GA 2002 USA

I. Chuckpaiwong S. A. Billington is with the Manufacturing Research Center at Georgia Tech, Atlanta, GA 2002 USA

A later version of the RVSM was developed for use in the 1996 Olympics held in Atlanta, Georgia and was addressed in a paper presented by one of the authors.⁸ This system was built to monitor the heartbeat of competitors in the archery and rifle events and was able to penetrate through the heavy leather flak jackets typically used by competitors. Finally, a variant called the RADAR Flashlight was developed for use by law enforcement personnel to detect the radar respiration signature of individuals concealed behind a wall or within an enclosed space under the sponsorship of the National Institute of Justice (NIJ).⁹ A picture of the latest Radar Flashlight prototype is shown in Figure 2.



Figure 2. Radar Flashlight prototype

Recent advances in the technology have increased the resolution of the sensor so it is able to detect motion on the order of microns. The associated hardware and signal processing advancements have now enabled the sensor to detect vibrational skin motion associated with speech directly off of the body.

III. SURFACE VIBRATION SPEECH SENSOR THEORY

The radar microphone is based on a phase detection technique to achieve a sensitivity high enough to pick up surface vibrations due to human speech. The key to the technique is that it does NOT use the Doppler effect or time of flight measurements common in most traditional radar designs. The key to the GTRI technique is that the sub-wavelength phase is measured with high accuracy. Motion less than the transmitted wavelength is being measured.

The radar microphone detects motion similar to a laser vibrometer, however, millimeter microwaves are used instead of light and a homodyne detection technique is being used instead of an interferometer. Typically, when electromagnetic waves are used in the context of radar or other remote sensing applications, the object of interest is moving through multiple wavelengths. If that object is moving relative to the transmitter, the received frequency will be different than the transmit frequency. This is the well-known Doppler effect. However, when an object moves less than a wavelength, such as the case in detecting chest vibrations, a different phenomenology, phase modulation, is at work.

To prove the basic fundamentals of the concept, the vibration of the chest was first recorded with a contact accelerometer and the corresponding acoustic speech was recorded with a microphone. The accelerometer was a high frequency PCB 352C68 placed on the chest and the

microphone was a standard acoustical transducer. The simultaneously recorded output from the two sensors for the segment of speech "hickory dickory dock" is shown in Figure 3. The accelerometer data clearly shows many of the same characteristics as the audio signal. The radar microphone will measure the same vibrations as the accelerometer in a non-contact manner. Past research by the authors has shown that signal detected by the radar correlates well with accelerometer outputs.¹⁰

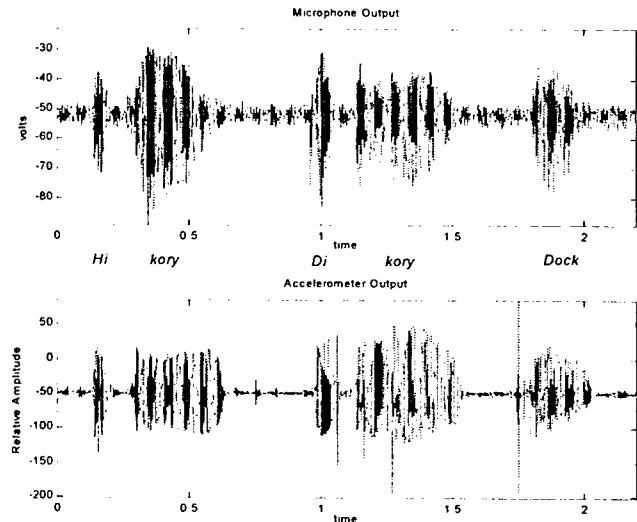


Figure 3. Simultaneous microphone and accelerometer speech data for "hickory dickory dock"

IV. PROTOTYPES

A prototype has been constructed to demonstrate the technology for a different application; however, the results are useful to show the current state of the technology as well as the promise of the radar microphone. The resulting hardware was tested using a linear motor with an optical encoder.

Figure 4 depicts the hardware configuration of the test setup. A target was attached tightly to a moving portion of a linear motor. The target surface was covered with a flat metal sheet that is used as a reflector. The radar sensor and the linear-motor encoder were set to take simultaneous measurements. The displacement from the radar sensor and the encoder were compared, consequently the radar sensor could be calibrated and compared.

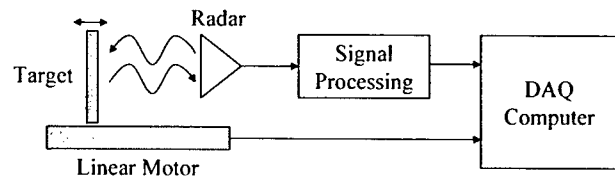


Figure 4. Radar microphone test setup

The results are illustrated in Figure 5. The top graph is a plot of both the radar sensed motion, and the ground truth motion as recorded by the encoder. It can be seen that the radar sensor was able to track actual displacement of an

arbitrary motion. The residual (difference between the radar and encoder calculated displacement) on the lower graph is the difference between displacements measured by the radar sensor and the reference, or error, of the radar sensor. According to this graph, the accuracy of the radar sensor can be given to within ± 1 mm over a displacement range of 50mm. Looking at smaller portions of the displacement, it can be seen that the error is often less than 0.1 mm.

Also, the residual being measured in this case is absolute displacement. Relative displacement errors have been measured down to 20 microns. Note that the residual is not randomly distributed, but a periodic function of displacement. The periodic error is caused by multipath reflections between the metal target and the metal radar hardware. Sensing of speech motion will yield significantly less multipath and distortion due to the less coherent reflecting surface.

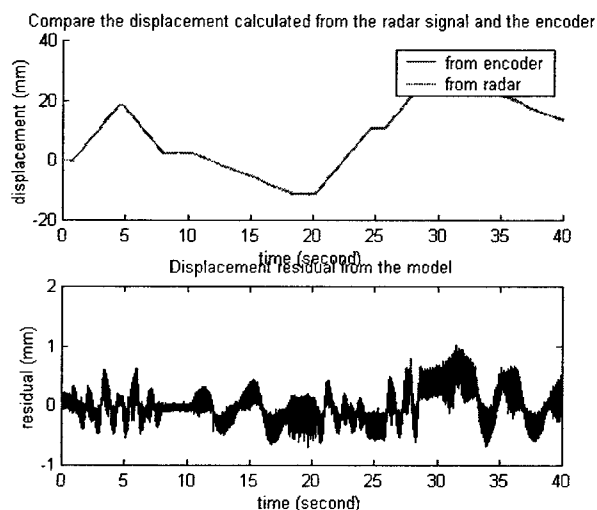


Figure 5. Example data taken from test setup

Some initial recordings have been taken using this prototype along with simultaneous acoustic recordings. After processing the radar signal, the presence of speech information is readily apparent at frequencies below 500 Hz and the signal correlates well with the acoustic data, however, the radar-derived speech is not yet intelligible. Increases in performance will occur both through signal processing, as well as better antenna design, which will increase the frequency response, as discussed below.

V. MODAL ANALYSIS

Critical to the successful operation of a radar microphone is the "spot size" of microwave energy illuminated by the antenna. This is critical because the sensor is measuring vibrations that are propagating along the surface of the chest. Waves with peaks and nulls are moving through the chest at different frequencies. One analogy would be the waves that move outward in water when a stone is dropped into a pond. There are peaks and nulls in the water corresponding to the propagating surface waves.

The work of Dr. Kevin Riggs at Stetson University has produced holographic images of vibratory modes in different materials. Figure 6 shows an example vibratory mode for a six inch square steel plate. The peaks and nulls on the plate are readily apparent. It is critical for accurate measurement of the vibration signal that the illumination area not detect both peaks and nulls at the same time, which may smear the output signal in the frequency domain.

Because the radar is receiving the sum of reflections from all illuminated points, the peaks and nulls could cancel each other out and distort the signal of interest. Therefore, the bandwidth of the radar microphone is limited by the antenna spot size on the chest. The smaller the spot size, the higher the frequencies that can be adequately picked up by the sensor.



Figure 6. Example image of vibratory modes on a steel plate (K. Riggs, Stetson University)

As the standoff distance from the radar to the target of interest increases, the area illuminated by the radar beam increases, affecting the frequency sensitivity of the sensor. The spot size in centimeters vs. distance in meters for various antenna beam sizes (in degrees) is shown in Figure 7.

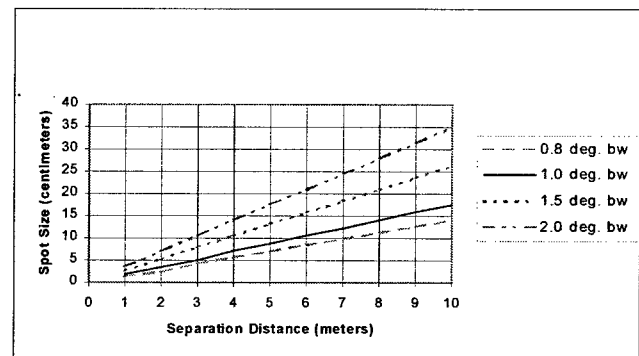


Figure 7. Spot size for given antenna beamwidths and distances

For the sensor to be viable, an antenna must be designed that projects a small spot size onto the neck, face, or chest of the person. If the application is in traditional military communications, the soldier or pilot will typically be wearing a headset, to which a sensor can be placed close to the face or neck. For larger standoffs, more exotic antennas will need to be designed. Moving the radar to a higher transmitted frequency will also enable smaller spot sizes, enhanced

resolution, and improved frequency response. As advances in commercial radar technology drive prices down for operating at higher frequencies (such as 77 GHz for automobile collision control), the ability of the technology to detect high resolution speech will be improved.

VI. CONCLUSION & FUTURE DIRECTIONS

The concept of using a radar device as a surface vibration electromagnetic speech sensor has been introduced. The radar acts as a sensitive motion detector able to detect the surface vibration of skin due to speech. Testing of a 35.0 GHz sensor has shown the ability to measure motion down to microns. The next step is to take the 35.0 GHz radar sensor and record a corpus of simultaneous radar and audio data to process and compare. Signal processing algorithms will be necessary to extract speech information out of the radar data. Initial recordings using the sensor have shown the presence of speech information at 500 Hz and below in the radar signal.

REFERENCES

- [1] Sundlberg, Johan, "Phonatory head and chest vibrations in singers," 127th ASA Meeting, June 6-10, 1994.
- [2] Viswanathan, V., Henry, C., "Noise-immune multisensor speech input: formal subjective testing in operational conditions," ICASSP-89, Glasgow, UK, March 1989.
- [3] Viswanathan, V., Henry, C., "Noise-immune multisensor speech transduction using multiple sensors," ICASSP-85, Tampa, FL, March 1985.
- [4] Burnett, G. C., et al., "The use of glottal electromagnetic micropower sensors (GEMS) in determining a voices excitation function," Proceedings of the 138th Meeting of the Acoustical Society of America, Columbus, Ohio., November 2, 1999.
- [5] Holzhricht, J.F., Ng, L. C., "Speech articulator and user gesture measurements using micropower, interferometer EM-Sensors," IEEE Instrumentation and Measurement Technology Conference, Budapest, Hungary, May 21-23, 2001.
- [6] J. Seals, S. R. Crowgey, and S. M. Sharpe, U.S. Patent Number 4958638, issued September 25, 1990.
- [7] J. Seals, S. R. Crowgey, and S. M. Sharpe, "An electromagnetic non-contact vital signs monitor," *SOUTHCAN '87 Conference Record*, 1987.
- [8] E. F. Greneker, "Radar sensing of heartbeat and respiration at a distance with security applications," *Proceedings of SPIE, Radar Sensor Technology II*, Vol. 3066, pp. 22-27, Orlando, Florida, April 1997.
- [9] E. F. Greneker, "Radar Sensing of Heartbeat and Respiration at a Distance with Security Applications," *Proceedings of SPIE, Radar Sensor Technology II*, Vol. 3066, Orlando, Florida, pp. 22-27, April 1997.
- [10] J. L. Geisheimer, E. F. Greneker, "Neural network applications of the radarcardiogram (RCG)," *SPIE AeroSense '99*, Orlando, Florida, April 1999.

Evaluation of ASR Sensors

Justin Taylor¹, Jason Heinrich², Jung H. Kim¹, Sung H. Yoon²

¹Department of Electrical and Computer Engineering

²Department of Computer Science

North Carolina A&T State University

Greensboro, NC 27411

Tel: (336) 334-7760 x 219 Fax: (336) 334-7244

E-mail: jt981532@ncat.edu or jh015778@ncat.edu

Abstract

This paper addresses the testing and analyzing of various microphones versus the Physiological Microphone (provided by Pete Fisher of the Army Research Laboratory) in different working conditions [1,2]. We explore different techniques and environments in which a user interfaces a selected ASR program. The testing of multiple microphones provided us with varied results based on environment. The software of choice for our research was Dragon Naturally Speaking 5.0.

1. Introduction

Automatic Speech Recognition systems enable users to operate their computer through the use of their voice. This advancement has benefited casual consumers, professionals and handicapped individuals alike. The development of a microphone allowing the user to move about freely and eliminate background noise has become necessary for practical use by professionals and consumers alike. Although significant progress has been made in ASR there are still limitations that must be taken into consideration. The technology that is on the market for consumers today, operates efficiently only under controlled conditions and through dictation, not conversation.

Factors to be considered in recognition accuracy:

- Environment (background noise, room size)
- Computer Hardware (CPU speed, RAM, soundcard)
- Amount of training with software
- Position of microphone
- Speaking style and clarity

- Microphone type
- Variability in the consumers speech (e.g., stress, colds)

These factors are considered to determine the most effective speech recognition procedure for each microphone based on environment.

2. System Descriptions

Our research was recorded based on the results provided by two test machines. The machines were both using Intel based processors.

System A

- Pentium III 0.5 GHz
- 256 Mb pc133 RAM
- Yamaha DS-XG Sound Card

System B

- Pentium IV 1.4 GHz
- 256 Mb RDRAM
- Sound Blaster Live! 5.1

System C

- Pentium IV 1.4 GHz
- 256 Mb RDRAM
- Sound Blaster Live! 5.1

System D

- Pentium IV 1.8 GHz
- 256 Mb RDRAM
- SoundBlaster Live! 5.1

The testing phase of the research continued through the use of four styles of microphones.

Microphone types:

- Telex H-551 Headset Microphone (Reference Mic.) (System B)
 - USB digital stereo headset
- Physiological Microphone (P-Mic)
 - Throat Microphone that detects vibration through skin and bone (System A)
- Telex M-60
 - Super-directional linear array microphone (System C)
- Telex M-40
 - Standard desktop microphone (System D)

Our findings were based on the aforementioned hardware combined with a predetermined method of testing. All computers exceeded the hardware requirements of Dragon Naturally Speaking v5.0. Through preliminary testing, we found all recognizer engines operated at the same speed when dictating. Therefore, microphones were arbitrarily assigned to each computer.

3. P-Mic Description

The Physiological Microphone is optimized for hands-free use. The microphone is designed to eliminate most background noise. It has its own power source, which is a 7.5-volt silver-oxide battery. Two of the microphones we used were a stationary desktop microphone (Telex M-40) and a super-directional linear array based microphone (Telex M-60). The P-Mic has a power switch allowing the user to pause in dictation without having to remove the microphone or stop the program. The Telex M-40 is lacking a power switch, which is inconvenient in ASR. Physically, the P-Mic does not resemble a typical microphone. The P-Mic is worn like a collar, and has a silicon contact sensor which is placed slightly to the left or right of the throat, due to the symmetrical nature of the throat. The P-Mic is small and lightweight. The width of the collar and diameter of the sensor is about 1 inch. With the P-Mic the user can move about freely and have both hands available. Traditional microphones used in ASR require that the user remain stationary, thus limiting productivity in the workplace. The P-Mic plugs into the "Line-In" jack on the sound card via a phono plug, whereas traditional microphones use the microphone jack.

4. Procedure for Microphone Testing

Testing was performed in a typical, quiet research laboratory environment. Our research lab's dimensions are 22' x 17'. The room is prone to little outside noise interference. A radio playing a recorded talk radio conversation at variable volumes was used to produce background noise. The recorded talk radio show was selected for consistency, allowing each microphone to be subject to the same interference. The simulated conversation source was emitted 10' behind the speaker.

Before testing we positioned four computers such that they could be tested simultaneously by one user. Each of the four microphones was assigned arbitrarily to a computer. We then performed the basic training required according to the Dragon Naturally Speaking documentation. Next a 400-word passage was dictated once while correcting and training all errors that occurred. The 400-word passage contained general vocabulary. After training, the Telex M-40 and Telex M-60 were attached to a microphone stand and positioned directly in front of the speaker. The user then attached the H-551 and the P-mic enabling all four microphones to be tested at the same time. The speaker tested each microphone with background noise set at; no additional noise, 60dB, 70dB, and 80dB respectively. The environment where we tested had an average of 50 dB of background noise. The quiet conditions were to facilitate the peak performance of each of the four microphones.

The speaker then started Dragon Naturally Speaking on all four computers. The speaker read the passage speaking at an average volume of 80dB. With the speaker speaking at 80 dB and noise at 50 dB, the difference of 30 dB provides an ideal speech-to-noise ratio for ASR. The speaker's volume was chosen to keep him from resisting the urge to compete with the added background noise, especially at the highest level of noise (80dB). This allowed the experiment to be performed at speech-to-noise ratios varying from excellent to very poor for speech recognition purposes. Each test was performed three times per sound level and the results were averaged. The dictated passages were printed and saved for analysis of mistakes made during dictation.

5. Results

The results for the four microphones tested are documented in the plot below. Results per microphone in each environment are the average of three test sessions, recording the accuracy rate. The equation we used was $[(\text{Errors} / \text{Total Words}) * 100 = \text{Percent Error}]$; then, $[100 - \text{Percent Error} = \text{Accuracy Rate}]$. Each capitalization error, period, paragraph indentation, etc. was counted as an error, and a wrong word or a skipped word was counted as one error. Therefore, Type I and Type II errors were counted as one error. Multiple word phrases recorded in error in the place of one word were counted as one error (example: user says, "comma" and program records, "come on", = one error).

Table 1 contains the results for the microphones tested at each level of background noise. The last column depicts the total percentage change from quiet conditions to 80dB background noise.

Table 1. (Performance in %)

Mic. Type	No Noise	60dB	70dB	80dB	Total Chg.
H551	99.0	98.5	96.5	89.75	9.25%
M-60	98.75	97.25	92.5	85.25	13.5%
M-40	95.5	94.25	87.5	81.75	13.75%
P-Mic	97.5	96.0	93.75	92.0	5.5%

The graph below illustrates that microphone performance was above 94% accuracy when speech-to-noise ratios were ideal. Notice that the steepest drop for the acoustic microphones occurred between 70 and 80 dB, whereas the slope of the P-Mic continues along a fairly straight line. The P-Mic never dropped more than 3% between increased levels of background noise.

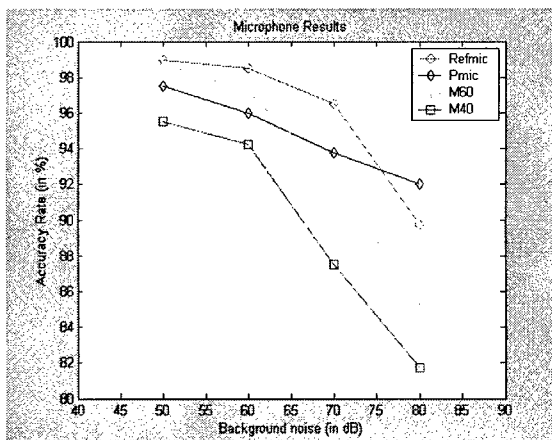


Figure 1 (Combined Results)

Table 2 breaks down the percent change in increased background noise. The acoustic microphones' performance all dropped in parallel as the levels of background noise were increased. The P-Mic's performance, on the other hand, did not decrease at a higher percentage with the addition of background noise. (Specifically from 60 to 70dB versus 70 to 80dB.

Table 2. (Percent Change)

Mic. Type	No Noise to 60dB	60 to 70dB	70 to 80dB
H551	0.5%	2.0%	5.25%
M-60	1.5%	4.75%	7.25%
M-40	1.25%	2.75%	5.75%
P-Mic	1.5%	2.25%	1.75%

5. Conclusions

It is concluded that the Physiological Microphone out performed its competition the most at the most stressful speech-to-noise ratios. The physiological microphone's performance was relatively unhampered by very poor speech-to-noise ratios. Our acoustic microphones' largest drop in recognition accuracy occurred at 80dB. The acoustic microphones dropped at least 5% at this level, whereas the P-Mic dropped only 1.75%. The P-Mic's total percent change of errors was about to half that of the reference microphone. Although the P-Mic performed above the rest, the 99% accuracy at quiet conditions still eluded it. Our data leads us to believe that the P-Mic has great potential when used in high background noise areas. We feel that the addition of an acoustic sensor used in tandem with the Physiological Microphone will boost recognition accuracy.

6. Future Endeavors

In the near future, we plan on acquiring a more accurate sound level meter, with a low range of 30dB. We would also like to acquire an electronic mouth to aid in our normalization process. Plans to create and implement a throat/neck simulator are also being arranged. This simulator, used with the electronic mouth will allow for a minimum of user errors and a near complete normalization of the test environment when using a pre-recorded file. We are also interested in acquiring other throat

sensors and testing their performance versus the Physiological Microphone.

7. References

[1] Pete Fisher: "Physiological Sensor For Speech Recognition", Proc. MultiModal Speech Recognition Worksop, Greensboro, NC (2002).

[2] Pete Fisher: "Alternative Speech Sensors For Military Applications", Proc. MultiModal Speech Recognition", Greensboro, NC (2002).

MULTI-MODAL SENSORY FUSION WITH APPLICATION TO AUDIO-VISUAL SPEECH RECOGNITION

Stephen M. Chu and Thomas S. Huang

Beckman Institute and Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

ABSTRACT

In this work we consider the bimodal fusion problem in audio-visual speech recognition. A novel sensory fusion architecture based on the coupled hidden Markov models (CHMMs) is presented. CHMMs are directed graphical models of stochastic processes and are a special type of dynamic Bayesian networks. The proposed fusion architecture allows us to address the statistical modeling and the fusion of audio-visual speech in a unified framework. Furthermore, the architecture is capable of capturing the asynchronous and temporal inter-modal dependencies between the two information channels. We describe a model transformation strategy to facilitate inference and learning in CHMMs. Results from audio-visual speech recognition experiments confirmed the superior capability of the proposed fusion architecture.

1. INTRODUCTION

Incorporating visual information into automatic speech recognition (ASR) has been demonstrated as an effective approach to improve the performance and robustness over the audio-only systems, and has received much attention in recent years [7]. One of the most challenging issues in bimodal ASR is how to fuse the audio (i.e. acoustic speech signal) and the visual (i.e. lip motion) modalities.

The fusion of audio and visual speech is an instance of the general sensory fusion problem. The sensory fusion problem arises in the situation when multiple channels carry complementary information about different components of a system. In the case of audio-visual speech, the two modalities manifest two aspects of the same underlying speech production process. From an observer's view, the audio channel and the visual channel represent two interacting stochastic processes. We seek a framework that can model the two individual processes as well as their dynamic interactions.

One interesting aspect of audio-visual speech is the inherent asynchrony between the audio and visual channels. Most *early integration* approaches to the fusion problem assume tight synchrony between the two. However, studies have shown that human perception of bimodal speech does not require rigid synchronization of the two modalities [6]. Furthermore, humans appear to use the audio-visual asynchronies as multimodal features. For example, it is well known that the voice onset time

(VOT) is an important cue to the voicing feature in stop consonants. This information can be conveyed bimodally by the interval between seeing the stop release and hearing the vocal cord vibration. Therefore, a successful fusion scheme should not only be tolerant to asynchrony between the audio and visual cues, but also be apt to capture and exploit this bimodal feature.

2. SENSORY FUSION USING CHMMs

It's a fundamental problem to model stochastic processes that have structure in time. A number of frameworks have been proposed to formulate problems of this kind. Among them is the hidden Markov model (HMM), which has found great success in the field of ASR. In recent years, a more general framework, the Dynamic Bayesian Networks (DBNs), has emerged as a powerful and flexible tool to model complex stochastic processes [3].

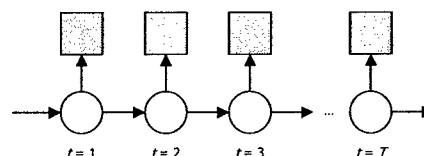


Figure 1. DBN representation of an HMM

The DBNs generalize the hidden Markov models by representing the hidden states as state variables, and allow the states to have complex interdependencies. Under the DBNs framework, the conventional HMM is just a special case with only one state variable in a time slice. DBNs are commonly depicted graphically in the form of probabilistic inference graphs. An HMM can be represented in this form by rolling out the state machine in time, as shown in Figure 1. Under this representation, each vertical slice represents a time step. The circular node in each slice is the multinomial state variable, and the square node in each slice represents the observation variable. The directed links signify conditional dependence between nodes.

It is possible to just use HMM to carry out the modeling and fusion of multiple information sources. This can be accomplished by attaching multiple observation variables to the state variable, and each observation variable corresponds to one of the information sources. Figure 2 illustrates the fusion of audio and visual information using this scheme. Because both channels share the single state variable, this approach in effect assumes the two information sources always evolves in lockstep. There-

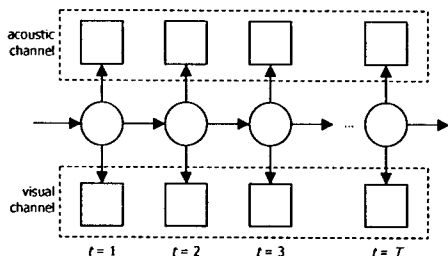


Figure 2. Audio-visual fusion using HMM

fore, it is not able to model asynchronies between the two channels.

An interesting instance of the DBNs is the so-called Coupled hidden Markov models (CHMMs). The name CHMMs comes from the fact that these networks can be viewed as parallel rolled-out HMM chains coupled through cross-time and cross-chain conditional probabilities. In the perspective of DBNs, an n -chain CHMM has n hidden nodes in a time slice, each connected to itself and its nearest neighbors in the next time slice. For the purpose of audio-visual speech modeling, we considered the case of $n=2$, or the 2-chain CHMMs. Figure 3 shows the inference graph of a 2-chain CHMM.

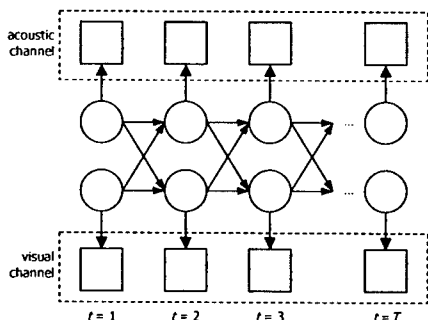


Figure 3. Audio-visual fusion using CHMM

There are two state variables in the graph. The state of the system at certain time slice is jointly determined by the states of these two multinomial variables. More importantly, the state of each state variable is dependent on both of its two parents in the previous time slice. This configuration essentially permits unsynchronized progression of the two chains, while encouraging the two sub-processes to assert temporal influence on each other's states. Note that the Markov property is not jettisoned by introducing the additional state variable and the directed links. Given the current state of the system, the future is conditionally independent of the past. Furthermore, given its two parents, a state variable is also conditionally independent of the other state variable.

In addition to the two state variables, there are two observation variables in each time slice. Each observation variable is a private child of one of the state variables. The observation vari-

ables can be either discrete or continuous. It is possible with this framework that one of the state variable is continuous and the other one is discrete.

In the context of audio-visual speech fusion, the audio and visual channels are associated with the two state variables respectively through the observable nodes. Inter-channel asynchrony is allowed. The overall dynamics of the audio-visual speech is determined by both modalities.

In general, the time complexity of exact inference in DBNs is exponential in the number of state variables per time slice. For systems with large number of state variables, exact inference quickly becomes computationally intractable. Consequently, much attention in the literature has been paid to approximation methods that aim to solve the general problem. Existing approaches include the *variational methods* [4] and the *sampling methods* [5]. However, these methods usually exhibit nice computational properties in an asymptotic sense. When the number of states is very small, the computational overhead embedded in the approximation method is often large enough to offset the theoretical reduction in time complexity. In this situation, the approximation becomes superfluous and exact inference becomes more desirable. In the following section, we describe a model transformation strategy that facilitates inference and learning in CHMMs.

3. CHMM TRANSFORMATION

The state of a 2-chain CHMM is jointly determined by the two state variables in the parallel chains. If the two state variables can take Q_1 and Q_2 discrete values respectively, then the CHMM in effect has $Q_1 \times Q_2$ possible states. The same state space can also be represented by a conventional HMM that has $Q_1 \times Q_2$ hidden states. Moreover, in CHMM, the output distribution of a joint state can be obtained by taking the product of the two output densities of the two individual state variables; Similarly, in a 2-stream HMM, the output distribution of a state is the product of the two stream-dependent densities. Hence, it is also possible to represent the output configurations of a 2-chain CHMM with a 2-stream HMM that has an equivalent state space. However, the observable nodes of a $Q_1 \times Q_2$ CHMM are fully specified by a table containing $Q_1 + Q_2$ entries. On the other hand, an unconstrained 2-stream HMM with $Q_1 \times Q_2$ hidden states has $2 \times Q_1 \times Q_2$ distinct output densities. This difference arises because in the CHMM an output node is only dependent on its single parent, while in the state-equivalent HMM the output is effectively conditioned on both state variables in the original CHMM. Fortunately, this discrepancy can be readily resolved through tying the appropriate output densities in the 2-stream HMM according to the mapping from CHMM states to HMM states. This state mapping and parameter tying procedure is easy to visualize graphically.

Figure 4 illustrates the state-machine diagram of 2-stream HMM obtained by transforming a 2-chain CHMM with $Q_1 = 3$ and $Q_2 = 2$. The state space of the original CHMM is repre-

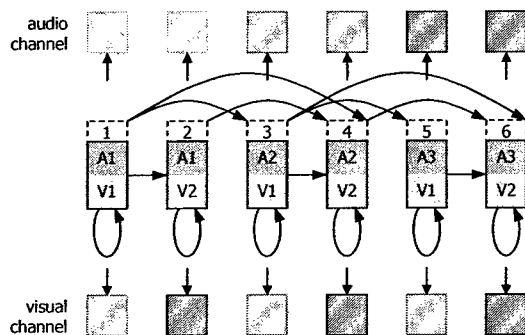


Figure 4. Transform CHMM to HMM through state-space mapping and parameter tying

sented by the 6 hidden states in the HMM. This mapping is explicitly depicted in the diagram. E.g., the state 3 in the HMM is equivalent to the state $\{q_1 = 2, q_2 = 1\}$ in the CHMM. The output densities of the HMM are tied according to the mapping. In the figure above, the observation nodes with the same color shade are tied. For example, the output densities modeling the lower stream in state 2, 4, and 6 are tied, because they all correspond to the entry $p(o_i | q_2 = 2)$ in the CPT of the CHMM.

The allowed state transition in the HMM is also derived from the state space mapping. In this example, it is assumed that the conditional probabilities concerning the two state variables in the CHMM satisfy the following condition.

$$P(q_{i+1}^1 | q_i^1, q_i^2) = 0 \text{ if } q_{i+1}^1 \neq q_i^1 \text{ and } q_{i+1}^2 \neq q_i^2 + 1 \quad (1)$$

This condition essentially enforces the left-to-right no-skip policy in the sense of conventional HMM for the two state variables in the CHMM, which is commonly used in audio-only speech recognizers. For example, a possible state path in the CHMM could be $\{q_1 = 1, q_2 = 1\} \rightarrow \{q_1 = 2, q_2 = 1\} \rightarrow \{q_1 = 3, q_2 = 2\}$, this is equivalent to the allowed state path $1 \rightarrow 3 \rightarrow 6$ in the HMM.

Other meaningful model configurations can be obtained through manipulating the allowed state transitions. For instance, it might be reasonable to model the dynamics of the lip motion using an ergodic state variable, i.e., no restriction on the possible state transitions for this variable.

It is worthy noting that the 2-stream HMM approach to audio-visual fusion as shown in Figure 2 can be considered as a special case of the CHMM-based fusion architecture. In that case, the number of the audio states must be equal to the number visual states, and the two state variables always progress in lock step, i.e. $Q_1 = Q_2$, and $q_1^t = q_2^t$ for all t . The CHMM-based fusion architecture permits a much richer space for modeling interactions between the two modalities.

The model transformation strategy described is fairly general and can be implemented on any HMM-based ASR platforms that support multiple observation streams and parameter tying.

4. AUDIO-VISUAL ASR EXPERIMENTS

The experiments carry two objectives. The first is to evaluate the improvement in noise robustness brought by the bimodal approach to ASR. The second is to compare the performance of the proposed fusion architecture with other fusion techniques.

To fulfill the first objective, we built an acoustic speech recognizer as the baseline system. The recognizer was trained using clean speech. Noisy condition of a particular SNR level was simulated by adding white Gaussian noise to the clean speech samples. The same acoustic feature sets were also used in the audio channel of the bimodal system. However, it is assumed that visual channel is not affected by any additional noise during testing. A visual-only recognizer was built and used as a benchmark. To achieve the second objective, we implemented a common form of the early integration approach, i.e. fusion by concatenating the audio and visual feature vectors. The systems were developed using HTK.

Evaluation of the bimodal speech recognition system was performed on an audio-visual speech dataset [1] collected by Chen *et al.* at the Carnegie Mellon University. The vocabulary consists of 78 words commonly used in scheduling applications. The visual features were derived from the lip-tracking data provided with the bimodal speech dataset. The primary visual features considered in the experiments are composed of h_1 , h_2 , which measure the vertical openings of the upper and lower lips, and the distance between the two mouth-corners, w . Delta features were also included, thus the actual visual feature vector is six-dimensional. The acoustic speech was processed using a 25ms Hamming window, with the frame period set at 10ms. For each frame, 12 MFCC coefficients were calculated from the result of filterbank analysis using 26 channels. Delta coefficients were also computed and then appended to the static features resulting in a 24-dimensional acoustic feature vector.

We constructed the acoustic and the audio-visual speech models at the word level. The audio-only system is based on HMMs with nine states, left-to-right topology, and no skips. The HMMs used in the visual-only system have a similar topology, but with only five states. HMM configuration identical to the audio-only system is used in the early integration bimodal system. The CHMM-based bimodal system uses five states to model the audio channel and three states for the visual channel. The allowed state transitions follow the policy specified in equation (1). Recognition was performed in the connected-word mode without the help of any grammatical constraints. A cross-validation scheme was used in the evaluations due to the limited amount of data. Specifically, the recognizers were trained on a subset containing 90% of the available data and tested on the remaining 10%; this process was repeated until all data had been covered in testing. The results are summarized in Table 1.

In the recognition results, it is evident that both of the bimodal systems demonstrate improved noise robustness in comparison to the audio-only system. However, at 10dB, the gain in robustness achieved by the early integration system is very lim-

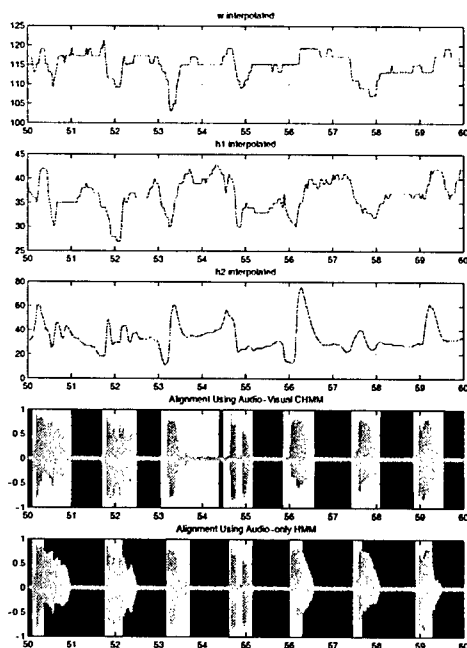


Figure 5. Forced alignment using audio only HMM and audio-visual CHMM

ited. On the other hand, the CHMM approach managed to give a clear improvement in performance at the same SNR level. At the 30dB, which is the SNR of the clean speech data, the recognition accuracy of the CHMM-based system is slightly worse than both the audio-only recognizer and the early integration bimodal system.

Table 1. Summary of recognition results (measured in %word accuracy). 'A' indicates the audio-only system; 'V' indicates the visual-only system; 'A+V' indicates the bimodal system using early integration; and 'CHMM' indicates the CHMM-based system.

SNR	10dB	20dB	30dB
A	4.03	43.61	99.10
V	42.95	42.95	42.95
A+V	10.58	72.79	99.74
CHMM	35.32	86.58	93.32

An important cue the visual modality provides in bimodal speech perception is the information about boundary locations of the speech units within an utterance. It would be interesting to see if this effect can be observed in our audio-visual ASR system. We computed forced alignment of a speech segment in the 20 dB test set using both the acoustic only recognizer and the CHMM-based bimodal recognizer. The results are illustrated in Figure 5.

Figure 5 covers a 10-second segment of the alignment result. The two subplots on the bottom show the word boundaries

superimposed with the speech waveform. The upper one is the alignment obtained using audio-visual CHMMs; the lower one shows the alignment obtained using acoustic only HMMs. The three subplots on the top display the static visual features used in the bimodal system. All five plots are time-aligned so that the correspondence among them can be visualized.

From the plot, we see that the audio-only recognizer almost always give the incorrect end-of-word boundary at this noise level. In contrast, the bimodal system was able to precisely determine the end boundaries in 6 out of 7 cases. It is interesting to observe that the bimodal recognizer consistently introduced a lead-time before the audible starting point of a word. This observation is consistent with the finding from human speech perception, that the visual speech usually leads the visual speech by a varying time window. The duration of the visual lead-in shown in Figure 5 ranges from about 40ms to 150ms.

5. CONCLUSIONS

We have described a novel sensory fusion architecture based on the CHMMs. A model transformation strategy that maps the state space of a CHMM onto the state space of a classic HMM is proposed to carry out inference and learning. Bimodal speech recognition experiments demonstrate that the CHMM-based fusion scheme can utilize the information in the visual channel effectively in noisy conditions.

6. REFERENCES

- [1] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, vol. 18(1), pp. 9-21, 2001.
- [2] S. Dupont and J. Luetin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2(3), pp. 141-150, 2000.
- [3] Z. Grahmani, "Learning dynamic Bayesian networks," in *Adaptive processing of temporal information* (C. L. Giles and M. Gori, eds.), Lecture notes in artificial intelligence, Springer-Verlag, 1997.
- [4] M. Jordan, Z. Grahmani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. I. Jordan, eds. Boston: Kluwer Academic Publishers, 1998.
- [5] D. J. C. Mackay, "Introduction to Monte Carlo methods," in *Learning in Graphical Models*, M. I. Jordan, eds. Boston: Kluwer Academic Publishers, 1998.
- [6] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, Cambridge: The MIT Press, 1998.
- [7] C. Neti, et al., *Audio-Visual Speech Recognition*, Final Workshop 2000 Report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, 2000.
- [8] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. IEEE ICASSP*, vol. 6, pp. 3733-3736, 1998.





Alternative Speech Sensors for Military Applications

U.S. Army Research Laboratory
Pete Fisher
6/10/02





Agenda

- Overview of current sensor technologies
- Possible future technologies
- Possible sensor fusion methods
- Military requirements
- Conclusion



Possible Methods for Improved ASR in Noise

- Reduce or eliminate noise through the processing of the speech signal based on properties of the signal components
- Detect speech without detecting noise
 - Many alternative sensors have reduced signal information
- Combinations of the above
 - Specialized sensors and processing
 - Multiple speech sensors and fusion of signals



Current Sensor Technologies

- Airborne acoustic microphones
- Contact acoustic microphones
- Bone conduction microphones
- Other alternative speech sensors



Airborne Acoustic Microphones

- Handheld microphones (Shure, etc)
- Headsets (Knowles, Shure, Telex, etc)
 - Noise canceling, close talking
- Super-directional microphones (Telex, etc)
 - Narrow band through beam forming
 - Linear arrays in a reinforcing pattern



Contact Acoustic Microphones

- Throat microphones (TEA, Genesys, Temco)
 - ARL Physiological Microphone
 - 32dB noise rejection
 - Acoustic response differs from a regular microphone
- Ear microphones (Jabra, Temco)
 - Some ear microphones are bone conduction
 - See next slide





Bone Conduction Microphones

- Navy bone conduction microphone
- Ear mounted bone conduction microphone
 - Invisio (TEA)
- Top of head bone conduction (Temco)
- Tooth mounted bone conduction microphone
 - Developed through a SBIR at CECOM





Other Alternative Speech Sensors

- Glottal Electromagnetic Micropower Sensor (GEMS)
 - Developed at Lawrence Livermore Nat. Labs
 - Commercial developer Aliph
 - Uses RADAR to measure internal motion
 - Reduced bandwidth
- Lip reading system (camera/computer)
 - Provides limited information, not a speech signal
 - Robust to noise



Other Alternative Speech Sensors

- Ultrasonic lip reader
 - Uses ultrasonic sensor to measure mouth opening
 - Have not been able to locate one of these devices, but have heard of them



Possible Future Technologies

- “Camera like” sensor that detects surface skin differently than tissues in the mouth
 - Would simplify detection of voiced speech
 - 3-5 and 8-12 micron FLIRs not suitable
 - Possibly some Near-Infrared technology?
- Novel vibration sensors
 - Technology?
 - Accelerometer? RADAR?



What to Sense?

- Vibration
 - Direct reading of speech or components
 - Close connection to avoid noise
- Alternatives?
 - Measure motion of speech articulators?
 - Tongue, teeth, glottis, sinuses
 - Modern jewelry?
 - Nose ring, cheek stud (microphones)
- Other methods?



Sensor Fusion

- Combining the outputs of one or more sensors to produce an improved speech signal
- Most appropriate in noisy environments where one or more sensors can be used to attempt to capture components of the speech signal while rejecting noise



Possible Sensor Fusion Methods

- Combine signals from multiple sensors in a cooperative fashion
 - Some non-standard speech sensors capture speech data while minimizing noise, but do not detect the full bandwidth of the speech signal
 - Could extract the cleanest spectral components of each sensor for input to ASR software



Possible Sensor Fusion Methods

- Use “clean speech” from noise robust sensors to remove noise from a primary sensor (airborne microphone)
 - Difference in secondary sensor signals and primary sensor signal is the noise (in the acoustic bands covered by the secondary sensors)
 - Could use correlation to remove noise that extends beyond the signal range of the sensor





Alternative Concept

- Work to improve a non-standard speech sensor and a matched ASR system to provide an integrated speech-in-noise package
 - Need a sensor with good noise rejection and “sufficient” signal capture capability
 - Need to tune the ASR engine to the peculiarities of the alternative speech sensor





Military Requirements

- Different for each application
 - Just like in the commercial world
- Selection of domain can be used to limit the problem
 - Command and control (C2) domain
 - Vocabulary of 1-5K words
 - Typically command phrases
 - Limited perplexity



Military Requirements (II)

- Most military environments will be noisy
 - Vehicles, people, weapons, generators, aircraft ...
- Capability to use existing microphones desirable in some cases
 - Communications via radio and vehicle intercoms
 - Difficulty of replacing all field equipment with improved or multi-modal speech sensors
 - Difficulty of getting more sensors on a soldier





What Do Military Users Want?

- They want a system that:
 - Works perfectly in all conditions
 - Weighs nothing
 - Is unbreakable
 - Does not interfere with their mission
 - Produces more energy than it uses
- Field soldiers are already overloaded
 - Make systems small (hand held), or make the software portable to platforms that are already carried by the soldier



Military Domains For SR

- C2 (command and control)
 - Constrained vocabulary, limited perplexity
 - “Tongue operated keyboard”
 - Electronic map navigation, radio settings
- Form completion
 - Repetitive task, limited vocabulary
 - Field reports, logistics (ordering supplies/ammo)
 - Might be performed over a low bandwidth field radio



Military Domains For SR (II)

- Information gathering
 - Vocabulary may not be constrained
 - User may have the option to enter free text fields with observations or other comments
 - Vehicle inspection, quality control
 - An actual military application of SR technology
- Monitoring of enemy communications
 - A much larger and more difficult application
 - Not amenable to application of alternative sensors



Conclusion

- There are a wide variety of alternative speech sensors available for exploitation for SR in military applications
- While many of these sensors do not detect the full range of human speech, their intrinsic noise rejection makes them useful
- Combinations of these alternative sensors may provide good solutions for the application of speech recognition in military environments

Session 2:
Audio-Visual Speech Recognition

Development and Evaluation of Audio-Visual ASR: A Study on Connected Digit Recognition

Michael T Chan
Rockwell Scientific Company
1049 Camino Dos Rios
Thousand Oaks, CA 91360
E-mail: mtchan@rWSC.com

Abstract

We present our findings from audio-visual speech recognition experiments for connected digit recognition in noisy environments. We derive hybrid (geometric- and appearance-based) visual lip features using a real-time lip tracking algorithm that we proposed previously. Using a small single-speaker corpus modeled after the TIDIGITS database, we build whole-word HMMs using both single-stream and 2-stream modeling strategies. For the 2-stream HMM method, we use stream-dependent weights to adjust the relative contributions of the two feature streams based on the acoustic SNR level. The 2-stream HMM consistently gave the lowest WER, with an error reduction of 83% at -3dB SNR level compared to the acoustic-only baseline. Visual-only ASR WER at 6.85% was also achieved. A real-time system prototype was developed for concept demonstration.

1. Introduction.

By combining acoustic and visual lip features for speech recognition, the resulting bimodal speech recognizer is markedly more robust in the presence of a variety of acoustic noise, when compared to the acoustic-only counterpart. The idea was pursued in a number of past studies [2][5][6][7][8][12][13][14][15][16][17][21]. Two key elements of an audio-visual speech recognition system are: (1) a front end for visual feature extraction, and (2) an information fusion architecture for integrating features from the two modalities. In recent years, considerable progress has been made in the first area [4][13][15][16], as well as in the second area [6][8][14][15][17].

There are primarily two categories of visual feature representation in the context of speech recognition. The

first is model-based or geometric-based. Examples of such features are the width and height of the mouth (and their temporal derivatives) that can be estimated from the images using a tracking procedure. The second category is pixel-based or appearance-based; that is, the features are directly derived from the raw pixel values. The first category is more intuitive, but there is typically a substantial loss of information because of the data reduction involved. There is little loss of information in the second representation, but the high dimensionality of the image space is a computational disadvantage, and pixel-based features do not directly relate to observable articulator motion. Furthermore, normalization needed to account for lighting changes, translation and other effects is more difficult compared to the geometric-based counterpart.

We had experimented with a visual feature representation that combined the two types of features in our previous work and demonstrated its effectiveness in simple isolated digit recognition experiments [4]. The technique is adopted in the work reported in this paper. Here we develop new experiments to evaluate our system using stream-weighted 2-stream Hidden Markov Models (HMMs) as well as the traditional single stream HMMs in the context of connected digit recognition.

The rest of the paper is organized as follows. We first briefly describe our lip localization and tracking algorithms that allow geometric-based features to be extracted automatically, and pixel-based features to be subsequently normalized. We then focus on the proposed hybrid feature and its efficacy in the context of visual-only speech recognition. Finally, we describe the recognition experiments we performed, and report our findings from these experiments involving audio-visual speech recognition of connected digits in the presence of aircraft cockpit noise of varying SNR levels.

2. Visual Tracking and Localization.

To automate machine lipreading, we need to locate and track movements and appearance changes of the lips. Several model-based approaches for tracking lip movements that have been proposed include snake models [10], deformable templates [20], active shape models [12], and active contours [11]. We have developed an integrated approach addressing both lip localization and lip tracking [2][3]. The first part is based on Gaussian mixture model-based clustering using hue in the HSV color space. The largest elliptical connected region detected with the expected range of hue values is identified as the lips. It is usually quite effective and can be used to initialize the lip tracking part. Tracking is based on a user-specific 2D B-spline model that can be constructed offline, or estimated from sample images [3]. To optimize tracking stability, the model deforms only in an affine subspace, which is adequate for capturing most lip movements that occur in normal speech utterances. The model is driven (or fitted) based on locations of steepest gradient in the image, in a linearly transformed color space given by

$$s = \alpha \cdot r + \beta \cdot g + \gamma \cdot b,$$

where $\{\alpha, \beta, \gamma\}$ are speaker-dependent and are estimated based on linear discriminant analysis on the RGB content [3]. This overcomes problems associated with often fuzzy definition of lip boundary in the luminance channel, and the algorithm is consequently markedly more robust compared to most snake-based algorithms and other approaches based on grayscale information alone. Another unique element is that the residual fitting error is used to monitor tracking errors and outlier measurements, and can trigger the lip localization module for automatic re-initialization. We have implemented a real-time tracking system on a 195MHz SGI O2 workstation that runs at 30fps. Figure 1 shows a few tracking examples.

3. Hybrid Visual Features.

Hybrid features are comprised of both geometric- and pixel-based features. Using tracking results obtained from the algorithm described above, geometric-based features, including the width and height of the mouth area and their temporal derivatives, can be estimated automatically. Pixel-based features are derived from the vertical intensity profile calculated based on a subset of the pixels, delimited by the boundary of the upper and lower lips explicitly estimated by the tracking algorithm. The number of pixels that defines the profile varies over time as the lips open and close. By proper sub-sampling and linear

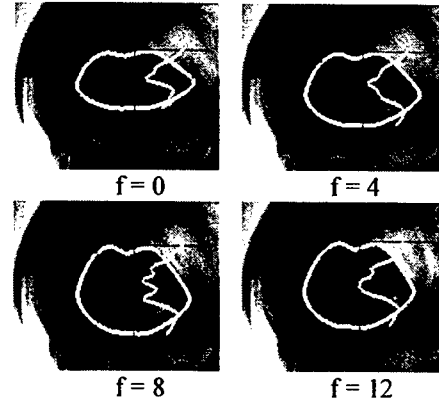


Figure 1: Snapshots of output from our lip tracking and visual feature extraction system in a few video frames. Geometric-based features were extracted from the tracking contour. Normalized pixel-based features were calculated based on the vertical intensity profile in the middle mouth region (plotted horizontally in light blue against a vertical axis).

interpolation, we map the vertical profile to a feature vector of constant length (e.g., 32 in our experiments). Therefore, information about the height of the mouth is largely decoupled from the pixel-based features. This is in contrast to cropping a rectangular region in the image that encompasses the lips in a sequence of image frames in an utterance, and subsequently taking the central vertical profile as the ROI. In practice, the ROI consists of a thin strip of pixels, where smoothing in the orthogonal direction is performed.

Robustness of ROI estimation for pixel-based features and the accuracy of tracking are known to be important for improving accuracy of visual speech recognition [9][13]. The approach we proposed could also be applied to the whole ROI defined by the tracking contour as opposed to only to the vertical profile. Furthermore, transform-based features similar to that in [15] could also be derived and used as features instead. Comparison with these variants will be a subject of future study. In our experiments, the center profile contained much of the information about the appearance of the teeth and tongue, as well as their spatial relationship, and good recognition accuracy was achievable even in visual-only speech recognition.

Figure 1 illustrates the application of the tracking algorithm for the extraction of visual features (both geometric- and pixel-based).

4. HMM for Audio-Visual Speech.

Here we describe the basic elements of the HMMs in our approach.

An N-state HMM is characterized by a state transition matrix, $\{a_{ij}\}$, $1 \leq i, j \leq N$, and a set continuous observation density functions, one for each state, which can be written as a Gaussian mixture

$$b_j(o_t) = \sum_{m=1}^M c_{jm} G(o_t; \mu_{jm}, V_{jm}), \quad 1 \leq j \leq N,$$

where o_t is the observation vector at time t , c_{jm} is the mixture coefficient, G is a multi-variate Gaussian distribution with mean μ_{jm} and covariance V_{jm} for m th mixture in the state j .

The acoustic and visual features were combined in two different ways in our HMM-based ASR experiments. In the first scheme, acoustic and visual feature vectors are concatenated to form individual feature vectors. In the second scheme, we model acoustic and visual features in separate feature streams. The mixture weights, mean vectors and covariance matrices in each observation density function are modeled separately in individual streams. The corresponding observation density is given by

$$b_i(o_t) = \left[\sum_{m=1}^{M_a} \alpha_{aim} G(o_{at}; \mu_{aim}, V_{aim}) \right]^{\beta_a} \left[\sum_{m=1}^{M_v} \alpha_{vim} G(o_{vt}; \mu_{vim}, V_{vim}) \right]^{\beta_v},$$

where subscripts a and v are used to denote the audio and visual channels, and the density of each channel is weighted by exponents β_a and β_v respectively, where $\beta_a + \beta_v = 1$. This is the multi-stream HMM formulation. The implicit assumption is that the audio and video observations are independent, which is really not exactly accurate. However, to be able to estimate reliably the parameters of b_i from limited amount of training data, it is customary to assume a diagonal covariance, and hence the assumption can be applied justifiably at least in the single Gaussian case with equal stream weights. Empirically, the stream weights can be used to give different emphasis to the observations, for example, based on the relative reliability of each channel.

5. Speech Recognition Experiments.

We performed a few evaluation experiments to compare various visual feature choices and investigate the relative merits of the various possible feature combinations. We focused on the connected digit recognition task. The

Table 1: Visual-only connected digit ASR's word error rate (WER %) for geometric (G), pixel-based (P), and hybrid (G+P) features described in this paper. The second and third rows are results with delta and delta-delta features. The size of the base feature vector is indicated in parentheses.

	G (2)	P(32)	G(2)+P(32)
Static	36.89	22.66	20.29
Static+ Δ	26.88	11.59	9.88
Static+ Δ + $\Delta\Delta$	27.80	9.49	6.85

eleven digits were 0-9 and 'oh.' The digit strings were taken from TIDIGITS, where utterances of up to seven digits were used. From a small database of 1518 audio-visual speech utterances, 759 were used for training and 759 for testing. Speech samples from one speaker were used to isolate the effects of speaker variability in this particular study. We used Hidden Markov Models to build word-model based recognizers. Gaussian mixtures were used to model the observation densities. The optimal number of mixtures (1-10) and number of hidden states (5-10) in the HMMs were determined empirically. A 3-state silence model was also used. The acoustic features were 12 Mel frequency cepstral coefficients (MFCC) plus the 0th order cepstral coefficient, as well as their first and second temporal derivatives, resulting in an acoustic feature vector of size 39. They were computed every 10ms using a 25ms frame analysis window. Per-utterance cepstral mean normalization was also applied.

The geometric features were derived from the width and height of the mouth normalized with respect to the corresponding dimensions when the speaker's mouth was closed. The pixel-based features were also normalized with respect to the mean value of the vertical profile when the speaker's mouth was closed. Interpolation of visual features was performed to generate samples at the audio feature frame rate of 100Hz.

In the audio-visual experiments, the audio features and visual features were concatenated to form a single feature vector for the single stream HMM case. The 2-stream HMM was also considered where the stream exponents were optimized using a linear step search. Alternatively, they could be discriminatively trained [17]. The Baum-Welch algorithm was used for EM-style embedded HMM training, and the Viterbi decoding algorithm for recognition. The HTK Toolkit [19] was used to design these experiments.

Table 1 shows first a summary of the recognition experiments employing visual features alone. One general trend we observed was that dynamic features (delta and delta-delta) in general carry additional information for

Table 2: Recognition WER (%) for the audio-only baseline (A), visual-only baseline (V), single stream audio-visual (AV1), 2-stream audio-visual (AV2) ASR at different SNR levels (dB). The reference visual feature used here was G+ Δ G+P. β_a is the optimal stream weight on the audio channel for AV2. Note that AV1 was worse than the visual-only ASR at -3dB, whereas AV2 remained better.

	clean	20	15	10	5	3	0	-3
A	0.13	0.66	5.53	23.58	67.19	75.63	80.11	85.11
V	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
AV1	0.13	0.53	1.32	2.50	7.38	10.14	15.55	22.79
AV2	0.13	0.26	0.53	2.50	6.59	9.75	12.12	14.49
β_a	0.95	0.85	0.8	0.65	0.5	0.45	0.35	0.35

recognition. Visual-only ASR word error rate as good as 6.85% was achieved, which was remarkable since no acoustic information was used and the pixel-based features were derived only from a small subset of pixels.

In the second experiment, we evaluated the effectiveness of the hybrid feature in the context of audio-visual speech recognition in the presence of noise. To be consistent with the visual features used in our previous work [4], the hybrid features employed were the combination of the base static pixel-based features, and the width and height of the mouth together with their first temporal derivatives (i.e., G+ Δ G+P). We added F-16 cockpit noise (from the NoiseX database) to the audio channel systematically at various SNR levels (20dB to -3dB) only to the testing data. Table 2 summarizes the results. We observe that the bimodal recognizers consistently outperformed the audio-only counterpart at all SNR levels. Furthermore, the 2-stream HMM outperformed the single-stream HMM, and the performance difference increased as the SNR decreased. That was possible because the 2-stream HMM allowed stream weights to be applied selectively based on reliability of the acoustic features. In fact, the optimal stream weight on the audio channel decreased monotonically with the SNR level. We expect the overall performance will be higher if we use all delta and delta-delta visual features.

Figure 2 shows a screenshot of the tracking and audio-visual ASR system prototype that we have developed for experimentation.

6. Conclusion.

We overviewed a real-time visual lip tracking system that we used to define the ROI for visual feature calculation.

We demonstrated the efficacy of our hybrid visual features in the context of connected digit recognition. Although single stream audio-visual HMM using concatenated features outperformed the acoustic-only counterpart, the 2-stream HMM gave the lowest WER at all SNR levels. The optimal stream weight for the audio channel decreased as the SNR level was lowered.

References

- [1] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *Proc. International Conference on Acoustics Speech and Signal Processing*, pp. 669-672, 1994.
- [2] M. T. Chan, Y. Zhang, and T. S. Huang, "Real-time lip tracking and bimodal continuous speech recognition," in *Proc. IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing*, pp. 65-70, 1998.
- [3] M. T. Chan, "Automatic lip model extraction for constrained contour-based tracking," in *Proc. IEEE International Conference on Image Processing*, Vol. 2, pp. 848-851, 1999.
- [4] M. T. Chan, "HMM-based audio-visual speech recognition integrating geometric- and appearance-based visual features," in *Proc. IEEE Workshop on Multimedia Signal Processing*, pp. 9-14, Cannes, France, Oct 3-5, 2001.
- [5] T. Chen, Rao, R. R., "Audio-visual integration in multimodal communication," in *Proceedings of the IEEE*, Vol. 86, pp. 837-852, 1998.
- [6] S. Chu, T. S. Huang, "Audio-visual speech modeling using coupled hidden Markov models," in *Proc. ICASSP*, 2002.
- [7] S. Gurbuz, Z. Tufekci, E. Patterson, J. Gowdy, "Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition," in *Proc. ICASSP*, 2002.
- [8] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary speech: looking ahead to practical speechreading systems," in D.G. Stork and M.E. Hennecke (eds.), *Speechreading by Humans and Machines: Models Systems and Applications*, Springer, 1995.
- [9] G. Iyengar, G. Potamianos, C. Neti, T. Faruque, and A. Verma, "Robust detection of visual ROI for automatic speechreading," *Proc. IEEE Workshop on Multimedia Signal Processing*, Cannes, 2001.
- [10] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active Contour Models," *International Journal of Computer Vision*, vol. 1, pp. 321-331, 1987.
- [11] R. Kaucic and A. Blake, "Accurate, Real-Time, Unadorned Lip Tracking," in *Proc. 6th International Conference on Computer Vision*, pp. 370-375, 1998.
- [12] J. Luetin, N.A. Thacker, and S.W. Beet, "Visual speech recognition using active shape models and hidden Markov models," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, v 2, pp. 817-820, 1996.
- [13] I. Matthews, G. Potamianos, C. Neti, J. Luetin, "A comparison of model and transform-based visual features for audio-visual LVCSR," in *Proc. International Conference on Multimedia Expo*, 2001.
- [14] S. Nakamura, K. Kumatani, S. Tamura, "Robust bi-modal speech recognition based on state synchronous modeling and stream weight optimization," in *Proc. ICASSP*, 2002.

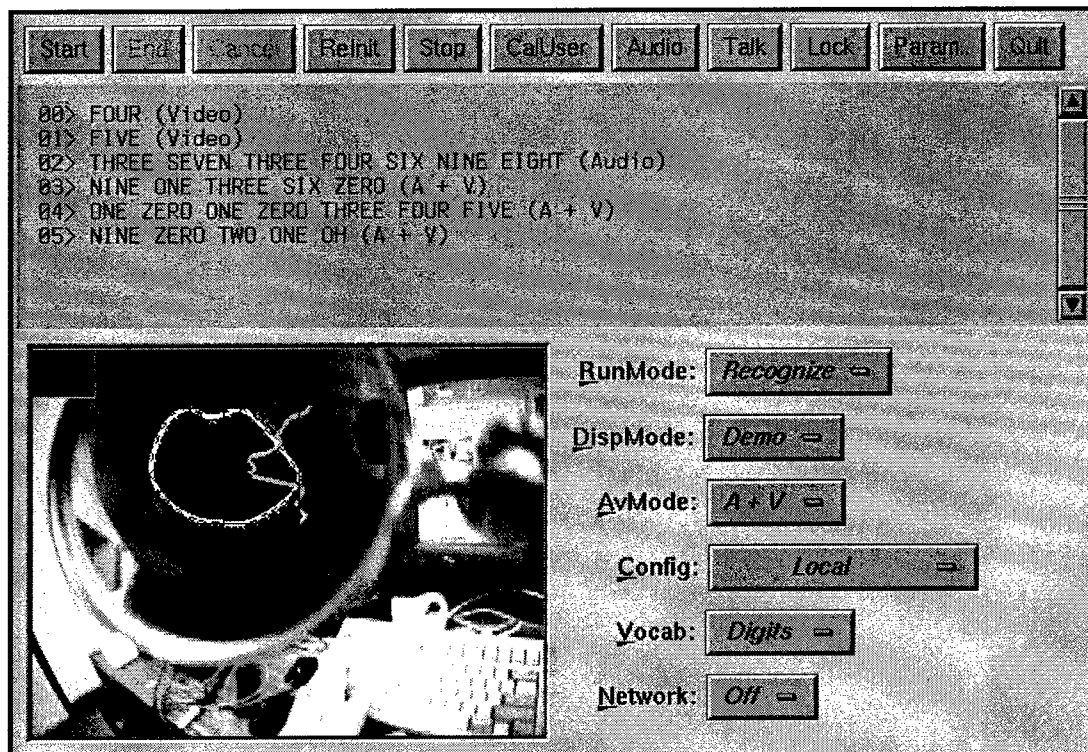


Figure 2: A screenshot of an experimental tracking and audio-visual ASR system at Rockwell Scientific. The system allows online switching among three recognition modes: audio-only, visual-only, or audio-visual. It can also be used to collect synchronized audio-visual sample data at 30fps directly to a disk array. A lightweight head-worn audio-visual capture apparatus can also be employed to allow users the freedom of head movement.

- [15] C. Neti, G. Potamianos, J. Luetttin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop," In *Proc. IEEE Workshop on Multimedia Signal Processing*, Cannes, 2001.
- [16] E. D. Petajan, B. Bischoff, and D. Bodoff, "An improved automatic lipreading system to enhance speech recognition," in *ACM SIGCHI-88*, pp. 19-25, 1988.
- [17] G. Potamianos, C. Neti, "Stream confidence estimation for audio-visual speech recognition," in *Proc. ICSLP*, vol III, pp. 746-749, 2000.
- [18] R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, New Jersey, 1993.
- [19] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK-Hidden Markov Model Toolkit V2.1*, Entropic Research, Cambridge, 1997.
- [20] A. L. Yuille, P. Hallinan, and D. S. Cohen, "Feature Extraction from Faces Using Deformable Templates," *International Journal of Computer Vision*, vol. 1, pp. 99-112, 1992.
- [21] Y. Zhang, S. Levinson, and T. Huang, "Speaker independent audio-visual speech recognition," in *Proc. International Conference on Multimedia and Expo*, Vol 2, pp. 1073-6, 2000.

Multimodal Dialog Systems Research at Illinois

Stephen E. Levinson, Thomas S. Huang, Mark A. Hasegawa-Johnson,
Ken Chen, Stephen Chu, Ashutosh Garg, Zhinian Jing,
Danfeng Li, John Lin, Mohamed Omar, and Zhen Wen

June 5, 2002

Abstract

Multimodal dialog systems research at the University of Illinois seeks to develop algorithms and systems capable of robustly extracting and adaptively combining information about the speech and gestures of a naive user in a noisy environment. This paper will review our recent work in seven fields related to multimodal semantic understanding of speech: audiovisual speech recognition, multimodal user state recognition, gesture recognition, face tracking, binaural hearing, noise-robust and high-performance acoustic feature design, and recognition of prosody.

1 Introduction

The purpose of this paper is to summarize ongoing multimodal speech and dialog recognition research at the University of Illinois. A multimodal speech recognition system can be described in two distinct stages: (1) robust audiovisual feature extraction, and (2) speech and user state recognition using dynamic Bayesian networks. Features are extracted from audiovisual input in order to optimally represent phonetic, visemic, gestural, and prosodic information. Our specific ongoing research projects include binaural hearing (array processing on a mobile platform), biomimetic noise-robust acoustic feature extraction, maximum mutual information acoustic feature design, and face tracking. Customized Dynamic Bayesian networks have been designed for three different recognition tasks: audiovisual speech recognition using coupled HMMs, user state recognition using hierarchical HMMs, and recognition of speaking rate using hidden-mode explicit-duration acoustic HMMs.

Image and Speech Processing research at the University of Illinois is currently tested in two ongoing research prototype environments. The first research prototype environment is an experimental computing facility for teaching children about physics. The sec-

ond research environment is an autonomous robot, *Illy*, who acquires language through the semantic association of audio, visual, and haptic sensory data. Prior to implementation on one or both of these platforms, most of our algorithms are tested using standard or locally acquired datasets.

2 Pre-Processing

2.1 Binaural Hearing

Our research on binaural hearing addresses the extraction of noise-robust audio from a two-microphone array mounted on a physically mobile platform (a language-learning autonomous robot). The source localization algorithm is based on a two channel Griffiths-Jim beamformer [3] and a new phase unwrapping algorithm for accurate estimation of time difference of arrival measures [8]. The new phase unwrapping algorithm is trained using many measurements of TDOAs in order to create an accurate spatial map of TDOA pattern as a function of arrival azimuth and elevation. These can then be used both to cancel interfering noise and to get a faithful representation of the desired speech signal. Preliminary results show that a speech signal can be accurately located in noisy laboratory room within a few milliseconds and with ten degree accuracy at a distance of 2-4 meters (acoustic far field).

In the current implementation, detection of a speech signal triggers physical rotation of the receiver platform (the robot's "head") so that it faces the primary talker. By physically aligning the "head" of the robot with the direction of primary source arrival, we are able to use extremely efficient off-axis cancellation algorithms for improved SNR [9].

2.2 Acoustic Features

Standard speech recognition features (including MFCC, PLP, and LPCC) result in isolated digit

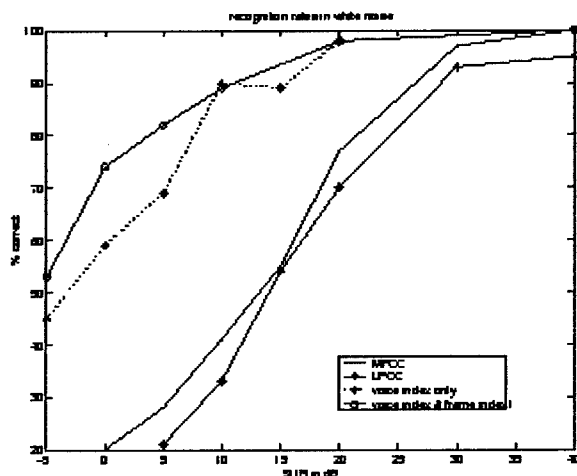


Figure 1: WER: isolated digit recognition in white noise with two standard feature sets, MFCC and LPCC, and two novel feature sets, LPCC with voice index and with frame index (from [6]).

recognition error rates of approximately 60% at 10dB SNR, and nearly 80% at 0dB SNR. In 1992, Meddis and Hewitt proposed a biomimetic method for recognition of voiced speech in high noise environments [10]. Meddis and Hewitt proposed filtering a noisy speech signal into many bands, computing the autocorrelation function $R_k(\tau)$ in each sub-band, and then estimating the speech autocorrelation $R(\tau)$ by optimally selecting and adding together the high-SNR sub-band autocorrelations. In our work [6], we have replaced Meddis and Hewitt's optimal selection algorithm by an optimal scaling algorithm. Specifically, we estimate the sub-band SNR v_k using a standard pitch prediction coefficient, i.e.

$$v_k = \frac{\text{Speech Energy in Band } k}{\text{Total Energy in Band } k} \approx \frac{R_k(T_0)}{R_k(0)} \quad (1)$$

where T_0 is the globally optimum pitch period. The maximum likelihood estimate of the noise-free speech signal autocorrelation is then

$$\hat{R}(\tau) = \sum_k v_k R_k(\tau) \quad (2)$$

In isolated digit recognition experiments, the use of equations 1 and 2 reduced word error rate by more than a factor of three in white noise at 10dB through -10dB, and by more than a factor of two in babble noise at the same SNRs (Figure 1).

The phonological features implemented at a speech landmark influence the acoustic spectrum at distances of 50-100ms [4, 19]. Complete representation of a 100ms spectrogram requires a 120-dimensional

Features	No LM		Phone Bigram	
	35dB	10dB	35dB	10dB
LPCC	56	40	59	46
MFCC	58	42	63	48
FM	58	42	62	46
MMIA	59	43	63	49

Table 1: Phoneme recognition correctness in four conditions. Features selected using a maximum mutual information criterion (MMIA) provide superior performance in all four conditions.

acoustic feature vector. It is not possible to accurately train observation PDFs of dimension 120 using existing data sets, but it is possible to select a sub-vector using a quantitative optimality criterion. In our research, we select a 39-dimensional feature sub-vector from a list of 160 candidate features in order to optimize the mutual information between features and phoneme labels [12]. Optimality is determined using a clean speech database (TIMIT) with no language model, but the resulting optimality generalizes. As shown in Table 1, the resulting MMIA (maximum mutual information acoustic) feature vector outperforms all standard feature vectors under at least three conditions: in quiet and at 10dB SNR, without a language model and with an optimized phoneme bigram. Larger improvements may be obtained by testing the 5-10 best feature vectors generated during the mutual information search. The best recognition accuracy, obtained using the feature set with second-best mutual information, was 62% with no language model in quiet conditions.

2.3 Face Tracking

Research has shown that facial and vocal-tract motions are highly correlated during speech production [20]. Speech recognition using both audio/visual features is shown to be more robust in noisy environments [5]. Analysis of non-rigid human facial motion is a key component for acquiring visual features for audio/visual speech recognition.

In the past several years, research in our group has led to a robust 3D facial motion tracking system [16]. A 3D non-rigid facial motion model is manually constructed based on piecewise Bezier volume deformation model (PBVD). It is used to constrain the noisy low-level optical flow. The tracking is done in a multi-resolution manner such that higher speed could be achieved. It runs at 5 fps on an SGI Onyx2 machine. This tracking algorithm has been successfully used for audio-visual speech recognition and bimodal emotion recognition.

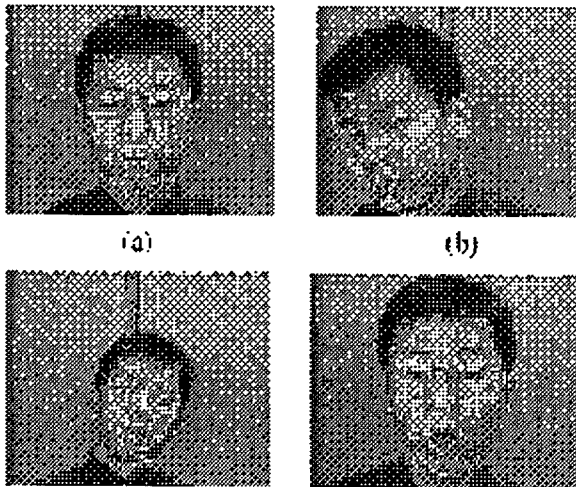


Figure 2: Demonstration of our face tracking system.

2.4 Gesture Recognition

Hand gestures are capable of delivering information not presented in speech [14]. Controlling gesture can be used to provide commands to the system. Navigation gestures provide information for manipulating virtual objects, and for selecting point objects or large regions on the screen. Conversational gestures provide subtle cues to sentence meaning in normal human interaction. Automated hand tracking and gesture recognition can help improve the performance of human-machine interface.

We have investigated both appearance-based gesture recognition (using neural network-based pattern recognition techniques) and model-based gesture recognition [18, 17]. In model-based recognition, the configuration of a hand model is first determined by providing a set of joint angle parameters. The 2D projection of this hand model, determined by the translation and orientation of the model relative to a viewing portal, is compared with the hand image from input video. Estimate of the correct input hand configuration is determined by the best matching projection. A complete description of the global hand position and all finger joint angles requires specification of 21 joint angles. Using both known anatomical constraints and PCA to reduce dimensionality, we can initially reduce the dimensionality of the gestural description from 21 to 7 independent dimensions while keeping 95% of the information. In this 7-dimensional space, it is possible to define 28 basis configurations, consisting of the configurations in which each finger is either fully folded or completely extended. A close examination of the motion trajectories between these basis states shows that natural hand articulations seem to lie largely in the linear

manifold spanned by pairs of basis states. We believe that, based on these preliminary results, it will be possible to map all observed gestures into a low-dimensional gestural manifold, resulting in efficient and accurate gesture recognition.

3 Dynamic Bayesian Networks

3.1 Lip Reading

The focus of our research in lip reading is a novel approach to the fusion problem in audio-visual speech processing and recognition. Our fusion algorithm is built upon the framework of coupled hidden Markov models (CHMMs). CHMMs are probabilistic inference graphs that have hidden Markov models (HMMs) as sub-graphs. Chains in the corresponding inference graph are coupled through matrices of conditional probabilities modeling temporal dependencies between their hidden state variables. The coupling probabilities are both cross chain and cross time. The latter is essential for capturing temporal influences between chains. In a bimodal speech recognition system, two-chain CHMMs are deployed, with one chain being associated with the acoustic observations, the other with the visual features. Under this framework, the fusion of the two modalities takes place during the classification stage. The particular topology of the CHMM ensures that the learning and classification are based on the audio and visual domains jointly, while allowing asynchronies between the two information channels.

In essence, CHMMs are directed graphical models of stochastic processes and are a special type of Dynamic Bayesian Networks (DBNs). The DBNs generalize the HMMs by representing the hidden states as state variables, and allow the states to have complex interdependencies. The DBN point of view facilitates the development of inference algorithms for the CHMMs. Specifically, two inference algorithms are proposed in this work. Both of the algorithms are exact methods. The first is an extension of the well-known forward-backward algorithm from the HMM literatures. The second is a strategy of converting CHMMs to mathematically equivalent HMMs, and carrying out learning in the transformed models.

The benefits of the proposed fusion scheme are confirmed by a series of preliminary experiments on audio-visual speech recognition. Visual features based on lip geometry are used in the experiments. Furthermore, comparing with an acoustic-only ASR system trained using only the audio channel of the same dataset, the bimodal system consistently demonstrates improved noise robustness across

SNR	10dB	20dB	30dB
A	4.03	43.61	99.10
V	42.95	42.95	42.95
A+V	10.58	72.79	99.74
CHMM	35.32	86.58	93.32

Table 2: Result of experiments in audiovisual speech recognition (measured in %word accuracy). A indicates the audio-only system; V indicates the visual-only system; A+V indicates a bimodal system using early integration; and CHMM indicates the CHMM-based system.

a wide range of SNR levels.

3.2 Prosody

Our approach to the recognition of prosody is the use of a "hidden mode variable" [13] to condition the explicit duration PDFs of a CVDHMM [7]. In our prototype algorithm, the state space consists of parallel phonetic state variables (q_t) and prosodic state variables (k_t). The dwell time of state q_t is a random variable d_q with PDF depending $p(d_q|q, k)$. At the end of the specified dwell time, the phonetic variable always changes state (no self-loops), but the prosodic state variable may or may not change state. Thus, for example, if ($k_t \in \text{slow, medium, fast}$) represents speaking rate, it may be reasonable to allow k_t to change state at any word boundary with a small probability.

In order to allow efficient experiments, we have modified HTK to make use of Ferguson's EM algorithm for explicit-duration HMMs [1, 2]. Ferguson's algorithm is an order of magnitude faster than most algorithms for the explicit-duration HMMs. The computational complexity of the algorithm is $O(NT(N+T))$, where N is the number of states, T is the number of frames in the input signal, and ($O(N^3T)$) is the complexity of an HMM without explicit duration. The forward algorithm computes

$$\begin{aligned}
\alpha_t^*(j) &= P(O_1, \dots, O_t, j \text{ commences at } t+1) \\
&= \sum_j \alpha_t(i) a_{ij} \\
\alpha_t(i) &= P(O_1, \dots, O_t, i \text{ ends at } t) \\
&= \sum_d \alpha_{t-d}^*(i) p(d|i) p(O_{t-d+1}, \dots, O_t|i)
\end{aligned}$$

3.3 User State Recognition

Integration of a large number of sources for the purpose of multimodal user-state recognition can be accomplished using a hierarchical dynamic Bayesian

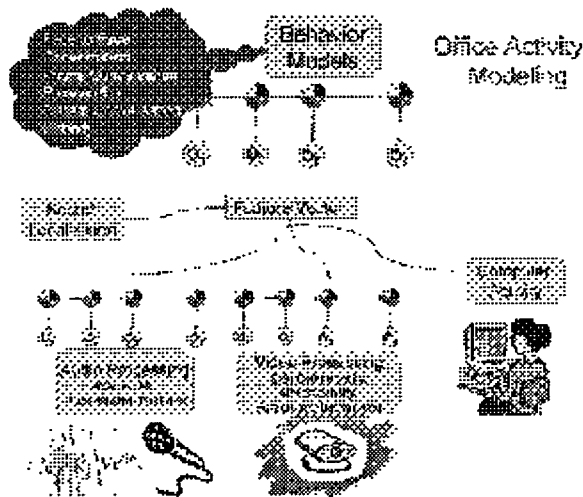


Figure 3: Architecture for detecting events in the office scenario

network (figure 3). In a hierarchical DBN, each modality (audio, lip reading, gesture, and prosody) is modeled using a modality-dependent HMM. Each modality-dependent HMM is searched in order to generate the N transcriptions that best match the observed data in the given modality. The likelihood of each transcription is then estimated using a constrained forward-backward algorithm, generating the probability of state residency during every frame. These probabilities are fed forward to the supervisor HMM, which integrates them to determine a single transcription of the sentence in order to maximize the a posteriori transcription probability. By imposing a prior on the probability distributions learned by the model for the purpose of increasing conditional entropy, we have demonstrated a 10% increase in user state classification performance [15, 11].

4 Conclusions

Our research is intended to elucidate both the theoretical and the practical requirements for effective multimodal speech understanding systems. The use of speech in multimodal systems will increase our theoretical understanding of the problems of sensor fusion and representations of multimodal signals. Increased theoretical understanding, in turn, will enable us to produce practical results that can be directly used in state-of-the-art speech recognition systems and as part of larger systems for advanced human-machine communication.

References

- [1] Ken Chen. Em algorithm for prosody-dependent speech recognition. Final Project Report, CS 346, 2002.
- [2] John D. Ferguson. Variable duration models for speech. In J.D. Ferguson, editor, *Proceedings of the Symposium on the Application of Hidden Markov Models to Text and Speech*, pages 143–179. Princeton University Press, Princeton, NJ, 1980.
- [3] L.J. Griffiths and C.W. Kim. An alternative approach to adaptive beamforming. *IEEE Trans. Antennas and Propagation*, AP-30(1):27–34, 1982.
- [4] Mark Hasegawa-Johnson. Time-frequency distribution of partial phonetic information measured using mutual information. In *Proc. Int. Conf. Spoken Lang. Proc.*, volume IV, pages 133–136, Beijing, 2000.
- [5] Marcus E. Hennecke, David G. Stork, and K. Venkatesh Prasad. Visionary speech: Looking ahead to practical speechreading systems. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines: Models, Systems, and Applications*, pages 331–350. Springer, New York, 1996.
- [6] Zhinian Jing and Mark Hasegawa-Johnson. Auditory-modeling inspired methods of feature extraction for robust automatic speech recognition. In *Proc ICASSP*, 2002.
- [7] Stephen E. Levinson. Continuously variable duration hidden Markov models for speech analysis. In *Proc. ICASSP*, pages 1241–1244, 1986.
- [8] Danfeng Li and Stephen E. Levinson. Adaptive sound source localization by two microphones. In *Proc. ICASSP*, page 143, Salt Lake City, UT, 2001.
- [9] Chen Liu, Bruce C. Wheeler, William D. O'Brien Jr., Robert C. Bilger, Charissa R. Lansing, and Albert S. Feng. Localization of multiple sound sources with two microphones. *J. Acoust. Soc. Am.*, 108(4):1888–1905, 2000.
- [10] Ray Meddis and Michael J. Hewitt. Modeling the identification of concurrent vowels with different fundamental frequencies. *J. Acoust. Soc. Am.*, 91(1):233–245, 1992.
- [11] Nuria Oliver and Ashutosh Garg. MIHMM: mutual information hidden markov models. In *Proceedings of the Nineteenth International Conference on Machine learning*, Sydney, 2002.
- [12] M. Kamal Omar and Mark Hasegawa-Johnson. Maximum mutual information based acoustic features representation of phonological features for speech recognition. In *Proc ICASSP*, 2002.
- [13] M. Ostendorf, B. Byrne, M. Fink, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In *CSLU Workshop 1996*, March 1997.
- [14] V. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gesture for human computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:677–695, 1997.
- [15] Vladimir Pavlovic, Ashutosh Garg, James M. Rehg, and Thomas S. Huang. Multimodal speaker detection using error feedback dynamic bayesian networks. In *IEEE Computer Vision and Pattern Recognition*, 2000.
- [16] H. Tao and T. Huang. Explanation-based facial motion tracking using a piecewise bezier volume deformation model. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1999.
- [17] Ying Wu and Thomas S. Huang. Self-supervised learning for visual tracking and recognition of human hand. In *Proc. AAAI National Conf. on Artificial Intelligence*, pages 243–248, 2000.
- [18] Ying Wu and Thomas S. Huang. View-independent recognition of hand postures. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 88–94, 2000.
- [19] Howard Yang, Sarel van Vuuren, and Hynek Hermansky. Relevancy of time-frequency features for phonetic classification measured by mutual information. In *Proc. ICASSP*, Phoenix, AZ, 1999.
- [20] Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26:23–43, 1998.

TEMPORAL ASYNCHRONICITY MODELING BY PRODUCT HMMS FOR AUDIO-VISUAL SPEECH RECOGNITION

Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories
satoshi.nakamura@atr.co.jp

ABSTRACT

There have been higher demands recently for Automatic Speech Recognition (ASR) systems able to operate robustly in acoustically noisy environments. This paper proposes a method to effectively integrate audio and visual information in audio-visual (bi-modal) ASR systems. Such integration inevitably necessitates modeling of the synchronization and asynchronization of the audio and visual information. To address the time lag and correlation problems in individual features between speech and lip movements, we introduce a type of integrated HMM modeling of audio-visual information based on a family of a product HMM. The proposed model can represent state synchronicity not only within a phoneme but also between phonemes. Furthermore, we also propose a rapid stream weight optimization based on GPD algorithm for noisy bi-modal speech recognition. Evaluation experiments show that the proposed method improves the recognition accuracy for noisy speech. In SNR=0dB our proposed method attained 16% higher performance compared to a product HMMs without the synchronicity re-estimation.

1. INTRODUCTION

The performance of ASR systems has been drastically improved recently. However, it is well known that the performance can be seriously degraded in acoustically noisy environments. Audio-visual ASR [1, 2, 4] systems offer the possibility of improving the conventional speech recognition performance by incorporating visual information, since the speech recognition performance is always degraded in acoustically noisy environments whereas visual information is not.

Audio and visual phonetic features have different durations. In other words, there is loose synchronicity between them, for instance, a speaker opens the mouth before making an utterance, and closes it after making the utterance. Furthermore, the time lag between the movement of the mouth and the voice might be dependent on the speaker or context.

As audio-visual integration methods for ASR systems, early integration and late integration are well known [1, 2]. In the early integration scheme, a conventional HMM is trained using audio-visual data. This method, however, cannot sufficiently represent the loose synchronization between the audio and visual information. Furthermore, the visual features of the conventional HMM may end up relatively poorly trained because of mis-alignments during the model estimation caused by the segmentation of the audio features. In the late integration scheme, the audio data and visual data are processed separately to build two independent HMMs

[1, 4]. This scheme assumes complete asynchronization between the audio and visual features. In addition, it can make the best use of the audio and visual data because there is a smaller bi-modal database than the typical database for audio only. However, the audio and visual features are regarded as independent. In this paper, in order to model the synchronization between audio and visual features, we propose pseudo-biphone product HMMs which realizes state synchronous audio-visual integration. The proposed model can represent synchronicity not only within a phoneme but also beyond phoneme boundaries. Furthermore, we propose a new method based on GPD algorithm to optimize stream weights of the proposed pseudo-biphone product HMMs.

2. AUDIO-VISUAL INTEGRATION BASED ON PRODUCT HMM

Figure 1 shows the outline of the acoustic model training for ASR systems in this paper. Figure 2 shows the proposed HMM topology. First, in order to create the audio and visual phoneme HMMs independently, audio features and visual features are extracted from audio data and visual data, respectively. In general, the frame rate of audio features is higher than that of visual features. Accordingly, the extracted visual features are incorporated such that the audio and visual features have the same frame rate. Second, the audio and visual features are modeled individually into two HMMs by the EM algorithm. Finally, an audio-visual phoneme HMM is composed as the product of these two HMMs based on HMM composition. The output probability at state ij of the audio-visual HMM is,

$$b_{ij}(O_t) = b_i^A(O_t^A)^{\alpha_A} \times b_j^V(O_t^V)^{\alpha_V} \quad (1)$$

which is defined as the product of the output probabilities of the audio and visual streams. Here, $b_i^A(O_t^A)^{\alpha_A}$ is the output probability

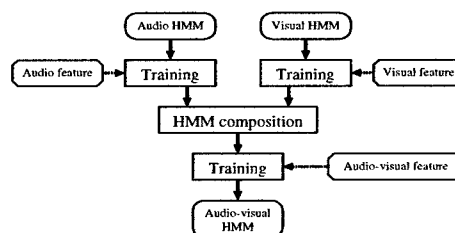


Fig. 1. Procedure Overview

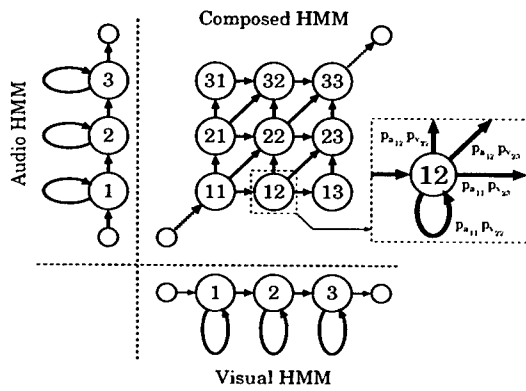


Fig. 2. Product HMM

of the audio feature vector at time instance t in state i , $b_j^V(O_t^V)^{\alpha_V}$ is the output probability of the visual feature vector at time instance t in state j , and α_A and α_V are the audio stream weight and visual stream weight, respectively. In a similar manner, the transition probability from state ij to state kl in the audio-visual HMM is defined as follows,

$$p_{ij.kl} = p_{a_{i.k}} \times p_{v_{j.l}} \quad (2)$$

where $p_{a_{i.k}}$ is the transition probability from state i to state k in the audio HMM, and $p_{v_{j.l}}$ is the transition probability from state j to state l in the visual HMM. This composition is performed for all phonemes. In the method proposed by [4], a similar composition is used for the audio and visual HMMs. However, because the audio and visual HMMs are trained individually, the dependencies between the audio and visual features are ignored. This results in the following two problems.

1. The product HMMs can not represent the loose synchronicity within phonemes as it is.
2. The product HMMs force a strict synchronization on every phoneme boundary.

This paper proposes a new approach to solve the two problems. The approach proposes re-estimation of the product HMMs parameters by using a small amount of audio-visual synchronous adaptation data, and pseudo-biphone product HMMs which represent loose state synchronicity beyond the phoneme boundary.

2.1. State Synchronous Modeling within a Phoneme

The first problem is from the inability of the conventional product HMMs to represent loose state synchronicity within a phoneme. This problem is caused by the fact that the transition probabilities and output probabilities are obtained by the multiplication of probabilities from independent states of audio and visual HMMs. We propose new product HMMs whose parameters are re-estimated using audio-visual synchronous adaptation data [3]. The re-estimation is able to introduce the loose state synchronicity of the states of two modalities into the product HMM. The re-estimation procedure is carried out using a small amount of audio-visual synchronous data. After the composition of two HMMs, the product HMMs can be re-estimated based on the Baum-Welch algorithm for multi-stream HMMs.

Figure 3 shows results comparing audio HMMs, visual HMMs, early integration, late integration, and product HMMs with and without re-estimation [3]. The experimental conditions are the same as those in a later section except that the audio HMMs are trained using clean speech data. The figure shows that the product HMMs with re-estimation achieve the best performance, while the product HMMs without re-estimation are worse than those of the early and late integration schemes.

2.2. State Synchronous Modeling Beyond The Phoneme Boundary

The second problem is that the conventional product HMMs force a strict synchronization on every phoneme boundary. This is because the speech organs normally move earlier than the speech to be produced. Sometimes, the speech organs are already articulated in the previous audio phoneme utterance. Accordingly, we have to consider state synchronous modeling beyond the phoneme boundary. We have carried out preliminary experiments using audio-visual word HMMs and confirmed that synchronicity is not always kept on a phoneme boundary looking at the optimal paths[5].

We propose new product HMMs that include extra asynchronous states on phoneme boundaries as indicated in Fig. 4. The core states of the phoneme HMMs are the same as those of context independent phoneme product HMMs. In addition, the new product HMMs have two extra HMM states aiming to work similarly to the word HMMs. The first extra state is composed of the initial audio state and final visual state of the preceding phoneme HMM. The second extra state is composed of the initial visual state and final audio state of the preceding phoneme HMM. Since these extra states are dependent on the preceding phoneme, they can only be re-estimated in a manner similar to the biphone HMMs. Therefore, we call these HMM pseudo-biphone product HMMs. The proposed HMMs can tolerate one state asynchronicity beyond a phoneme boundary.

3. STREAM WEIGHT OPTIMIZATION

As methods for estimating stream weights, maximum likelihood [6] based methods or GPD (Generalized Probabilistic Descent)[7] based methods have been proposed. However, the former methods have a serious estimation drawback because the scales of two probability are normally very different and so the weights can not be estimated optimally. The latter methods have substantial possibility for optimizing the weights. However, a serious problem is that these methods require a lot of adaptation data is necessary

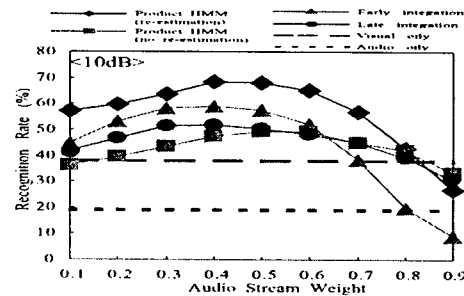


Fig. 3. Results of Product HMMs

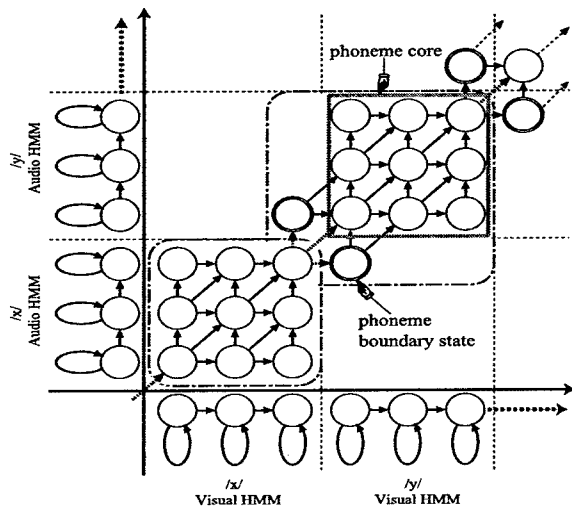


Fig. 4. Pseudo-biphone product HMMs

for the weight estimation. In this paper, we propose a GPD-based simplified adaptive estimation of stream weights using GMMs for new noisy acoustic conditions.

The approach by the GPD training defines a misclassification measure, which provides distance information concerning the correct class and all other competing classes. The misclassification measure is formulated as a smoothed loss function. This loss function is minimized by the GPD algorithm. Here, let $L_c^{(x)}(\Lambda)$ be the log-likelihood score in recognizing input data x for adaptation using the correct word model, where $\Lambda = \{\lambda_A, \lambda_V\}$.

In a similar way, let $L_n^{(x)}(\Lambda)$ be the score in recognizing data x using the n -th best candidate among the mistaken word models.

The misclassification measure is defined as,

$$d^{(x)} = -L_c^{(x)}(\Lambda) + \log\left[\frac{1}{N} \sum_{n=1}^N \exp\{\eta L_n^{(x)}(\Lambda)\}\right]^{\frac{1}{\eta}} \quad (3)$$

where η is a positive number, and N is the total number of candidates. The smoothed loss function for each data is defined as,

$$l^{(x)} = [1 + \exp\{-\alpha d^{(x)}(\Lambda)\}]^{-1} \quad (4)$$

where α is a positive number. In order to stabilize the gradient, the loss function for the entire data is defined as,

$$l(\Lambda) = \sum_{x=1}^X l^{(x)}(\Lambda) \quad (5)$$

where X is the total amount of data. The minimization of the loss function expressed by equation (5) is directly linked to the minimization of the error. The GPD algorithm adjusts the stream weights recursively according to,

$$\Lambda_{k+1} = \Lambda_k - \varepsilon_k E_k \nabla l(\Lambda), k = 1, \dots, \quad (6)$$

where $\varepsilon_k > 0$, $\sum_{k=1}^{\infty} \varepsilon_k = \infty$, $\sum_{k=1}^{\infty} \varepsilon_k^2 < \infty$, and E is a unit matrix.

In this paper, we propose to use GMMs instead of HMMs to find optimal stream weights not for the recognition. GPD training on GMMs is quite simple and requires smaller amount of training data. We use 18 mixture Gaussians for GMMs and train them using all of the training data.

4. EVALUATION EXPERIMENTS

The audio signal is sampled at 12 kHz (down-sampled) and analyzed with a frame length of 32 msec every 8 msec. The audio features are 16-dimensional MFCC and 16-dimensional delta MFCC. On the other hand, the visual image signal is sampled at 30 Hz with 256 gray scale levels from RGB. Then, the image level and location are normalized by a histogram and template matching. Next, the normalized images are analyzed by two-dimensional FFT to extract 6x6 log power 2-D spectra for audio-visual ASR. Finally, 35-dimensional 2D log power spectra and their delta features are extracted. For each modality, the basic coefficients and the delta coefficients are collectively merged into one stream. Since the frame rate of the video images is 1/30, we insert the same images so as to synchronize the face image frame rate to the audio speech frame rate. For the HMMs, we use a two-mixture Gaussian distribution and assign three states for the audio stream and two states for the visual stream in the late integration HMMs and the baseline product HMMs. In this research, we perform word recognition evaluations using a bi-modal database [1]. We use 4740 words for HMM training and two sets of 200 words for testing. These 200 words are different from the words used in the training. We perform experiments using 15, 25, and 50 words. The context of the data for the adaptation differs from that of the test data. In order to examine in more detail the estimation accuracy in the case of less adaptation data, we carry out recognition experiments using three sets of data, each as different as possible from the context. The size of the vocabulary in the dictionary is 500 words during the recognition of the adaptation data. The GPD algorithm convergence pattern is known to greatly depend on the choice of parameters. Accordingly, we set $N = 1$ in (3), $N = 0.1$ in (4), $N = 100/k$, and the maximum iteration count = 8.

We compared the processed product HMMs without re-estimation (Product-HMM(W/O Re-est.)), the proposed product HMMs with re-estimation (Product-HMM(W Re-est.)), the proposed pseudo-biphone product HMMs without re-estimation (Pseudo-Biphon(W/O Re-est.)), the proposed pseudo-biphone product HMMs with re-estimation (Pseudo-Biphon(W Re-est.)), and GMM for GPD-based stream weight optimization for acoustic SNR=15, 0, and -5dB. White noise was used to reduce the acoustic SNR in this experiment. The audio HMMs were trained using the SNR=15dB data. The results indicate that the re-estimation of the product HMMs is quite effective to improve the performance. The re-estimation is able to introduce the loose state synchronicity of the states of two modalities into the product HMMs. The state synchronous modeling beyond the phoneme boundary by a pseudo-biphone product HMM also results in significant improvements to the product HMMs. It is also confirmed that the re-estimation further improves performance of pseudo-biphone product HMMs. The figures show optimal stream weights for the maximum performance vary according to each method and acoustic SNR. The solid arrows show the results by simplified GPD-based stream weight estimation using 25 adaptation words. The proposed GPD-based simplified stream weight optimization algorithm successfully estimated stream weight with almost the best performance. In the SNR=-5dB environment, the estimated weight is not the optimal one. Figure 8 shows standard deviation of the word accuracy over various SNRs, a number of adaptation words, and a number of candidates in GPD training. It is confirmed the standard deviation in SNR=-5dB is bigger than the others and smaller number of adaptation words gives bigger standard deviations. In SNR=0dB our

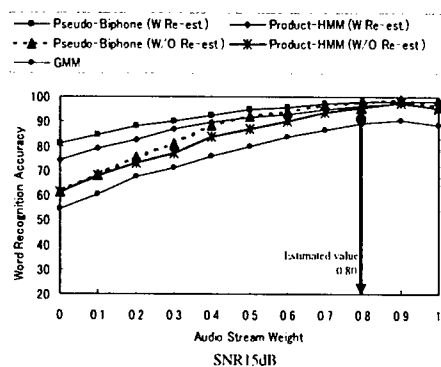


Fig. 5. Word Accuracy (SNR=15dB)

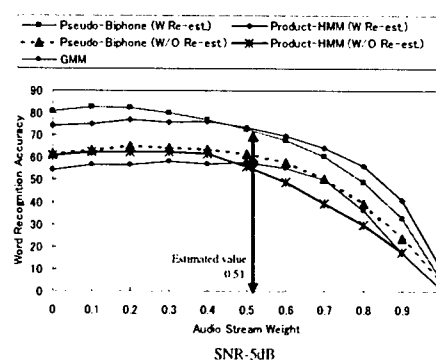


Fig. 7. Word Accuracy (SNR=-5dB)

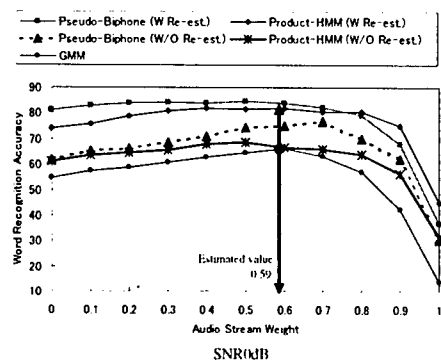


Fig. 6. Word Accuracy (SNR=0dB)

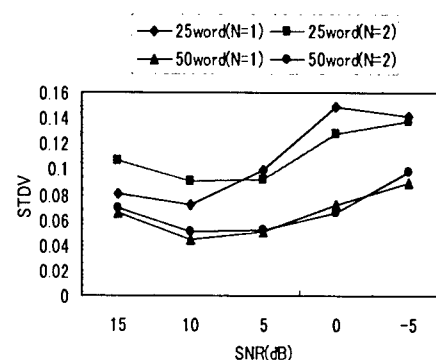


Fig. 8. Standard Deviation of Word Accuracy

proposed method attained 16% higher performance compared to a product HMMs without the synchronicity re-estimation.

5. CONCLUSION

This paper proposes a new HMM structure to effectively integrate audio and visual information in audio-visual (bi-modal) systems. Our state synchronous modeling of audio-visual information is based on the product HMM. The proposed model can represent synchronicity not only within a phoneme but also between phonemes. Evaluation experiments show that the re-estimation of the model parameters using audio-visual synchronous data further improves the product HMMs. In addition, pseudo-biphone HMMs that introduce two extra asynchronous states are shown to improve the bimodal speech recognition accuracy. Furthermore, we also proposed a rapid stream weight optimization based on GPD algorithm for noisy bi-modal speech recognition.

6. ACKNOWLEDGEMENTS

The authors thank intern students, K.Kumatani and S.Tamura, and their supervisors, Prof. S. Furui of the Tokyo Institute of Technology and Prof. K. Shikano of the Nara Institute of Science and Technology for giving us the opportunity to conduct this study col-

laboratively. This research was supported in part by the Telecommunications Advancement Organization of Japan.

7. REFERENCES

- [1] S.Nakamura, et al., "Improved bimodal speech recognition using tied-mixture HMMs and 5000 word Audio-Visual Synchronous database", Proc. Eurospeech97
- [2] S.Nakamura, et al., "Stream weight optimization of speech and lip image sequence for Audio-Visual speech recognition", Proc. ICSLP2000
- [3] Kenichi Kumatani, Satoshi Nakamura and Kiyohiro Shikano, "An Adaptive Integration Method Based on Product HMM for Bi-Modal Speech Recognition", HSC2001 (International Workshop on Hands-Free Speech Communication) pp. 195-198
- [4] M.J. Tomlinson, et al., "Integrating audio and visual information to provide highly robust speech recognition", Proc. ICASSP-96
- [5] S.Nakamura, K.Kumatani, S.Tamura, "State Synchronous Modeling of Audio-Visual Information for Bi-modal Speech Recognition", Proc. IEEE ASRU Dec. 2002
- [6] J.Hernando, "Maximum Likelihood Weighting of Dynamic Speech Features for CDHMM Speech Recognition", Proc. ICASSP'97,(1997) 1267-1270
- [7] G.Potamianos, H.P.Graf, "Discriminative Training of HMM Stream Exponents for Audio-visual Speech Recognition", Proc. ICASSP'98,(1998) 3733-3736

Visual Speech Feature Extraction From Natural Speech for Multi-modal ASR

Sabri Gurbuz and John N. Gowdy

Department of Electrical and Computer Engineering
Clemson University
Clemson, SC 29634, USA

E-mail:{sabrig,jgowdy}@ces.clemson.edu

Abstract

Improving the accuracy of speech recognition technology by addition of visual information is the key approach to multi-modal ASR research. In this work, we address two important issues, which are lip tracking and the visual speech feature extraction algorithm. In order to utilize the multi-modal ASR for natural speech, the visual front end algorithm must extract affine and lighting condition invariant visual speech features.

This paper focuses on both the lip tracking algorithm using the Bayesian framework and a novel pixel based visual speech feature extraction algorithm based on kurtosis measures of the frequency profile of the local image blocks. We compare the results of the proposed features with the results of outer lip contour based affine-invariant visual features, and global 2D DCT features. Experimental results in this paper are presented for a visual-only connected digit recognition task for performance comparison of the visual features.

Keywords: Lip tracking, Visual feature extraction, Kurtosis measure.

1. Introduction

The addition of visual information to audio features improves speech understanding and offers key advantages in human-computer interfaces especially in difficult environments [1–6]. Improving the existing state-of-the-art automatic speech recognition (ASR) performance by integrating the visual information of the speaker's mouth region is receiving significant attention from the speech recognition communities.

Some of the initial difficulties difficulty associated with computer lipreading (visual speech recognition) are the accurate and consistent visual region of interest (ROI) extraction, and lip tracking algorithm on the fly, which needs to be robust to a speaker's ethnic and gender variability, and other visual appearances such as glasses, facial hair, various skin color, lip color, and different lip shapes. Another difficulty difficulty is the robust and consistent visual speech feature extraction.

The development of a successful audio-visual speech recognition technology capable of adapting itself to changing environments will support both industrial and military applications. Audio-visual speech recognition research is a relatively new and advancing research area. A noise robust audio-visual speech recognition system will facilitate use of computers, increase reliability and worker productivity, and naturalize communications between human and computers. In addition, audio-visual speech recognition technology can facilitate new commercial applications such as

text-driven audio-visual talking head, audio-visual speech-to-speech translation, and speech-to-video conversion for the hearing impaired.

In our earlier research [1,7], we have implemented both late integration and early (multi-stream state synchronous) integration schemes for a controlled audio-visual data set. For both integration schemes, the experimental results showed that addition of visual information improves the recognition performance. In this paper, the following objectives will be sought:

1. Development of a lip tracking algorithm, and
2. A novel visual speech feature extraction algorithm that satisfies the following three criteria:
 - i. Affine (rotation, scale, and shear) invariance,
 - ii. Chrominance space shift invariance, and
 - iii. Chrominance space scale invariance.

In our proposed visual speech feature extraction method, the criteria in step (i) is satisfied by affine correction, the criteria in step (ii) is satisfied by removing of the DC component of the 2D DCT coefficients, and the criteria in step (iii) is satisfied by the normalized higher order moments of the DCT coefficients of the lip image blocks.

This work is organized as follows. In section 2, we present a Bayesian framework for lip tracking, parametric formulation of the Gaussian parameters and adaptation of the parameters on the fly. Section 3 discusses the removal of affine (rotation, scale, shear) effects from the segmented lip image. In section 4, we discuss contour based affine invariant features, pixel based normalized 2D DCT features, and describe a novel visual speech feature extraction algorithm based on kurtosis measures of the frequency profile of the local image blocks of the mouth. We present the experimental setup and the results in Section 5. Section 6 gives the concluding remarks and the proposed future work.

2. Lip Tracking Using the Bayesian Framework

The basis of the audio-visual speech recognition system is an efficient lip tracking algorithm. Computational time constraints required by applications such as audio-visual speech recognition, animated talking head design, etc., contribute to the difficulty of the task. Most lip tracking algorithms build upon the eigenspace based face detector and an ensemble of feature detectors which are used to extract pre-specified landmarks such as nostrils and lip corners to

locate the ROI (mouth region) [8, 9]. The deformable template and snake based methods [10, 11] have also been used for this task. All techniques have reported good results, but accuracy has decreased when there are occlusion (profile view), lighting condition change, texture changes, and quick motion. The technique we propose uses color images with Bayesian framework for classification which requires the estimation of the *a priori* probabilities and class conditional density models. The class conditional density and *a priori* probability estimation processes are described in the following sections.

In the lip tracking problem there are two distinct classes, *lip* and *non-lip*. Therefore, in this section, the two class classification problem is discussed because each sample in the image frame either belongs to *lip* class, w_1 or *non-lip* class, w_2 . The conditional density functions and the *a priori* probabilities are estimated using the training data that may require extensive search to locate the *lip* and *non-lip* regions in the first frame in practice which will not be discussed here. The Bayes decision rule determines whether an observation, x , belongs to w_1 or w_2 . One of the most commonly utilized probability density functions in practice is the Gaussian density function due to its computational simplicity and because it models a large number of cases in nature. The Gaussian parameters are estimated parametrically using the information from the previous frame on the fly which leads to an adaptive real time lip tracking and segmentation algorithm.

2.1. Parametric Formulation of Gaussian Density from Sample Data

In the parametric formulation of the multivariate Gaussian density, estimation of the mean vector and covariance matrices of the two classes, w_1 and w_2 , are required. Let N be the number of samples drawn from a class, w_i , with respect to x in the n -dimensional feature space. Then the general multivariate Gaussian (normal) density given by

$$p(x|w_i) = \frac{1}{\sqrt{(2\pi)^n \|\Sigma_i\|}} \exp\left\{-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right\}, \quad (1)$$

$i = w_1, w_2.$

where $\mu_i = E[x]$ is the mean value of the class w_i , and Σ_i is the $n \times n$ covariance matrices defined as

$$\Sigma_i = E[(x - \mu_i)(x - \mu_i)^T] \quad (2)$$

$\|\Sigma_i\|$ represents the determinant of Σ_i and $E[\cdot]$ is the expected value of a random variable. The parameters μ_i and Σ_i can be estimated without bias by the sample mean and sample covariance matrix as

$$\hat{\mu}_i = \frac{1}{N} \sum_{j=1}^N x_j^{(i)}, \quad i = w_1, w_2 \quad (3)$$

$$\hat{\Sigma}_i = \frac{1}{N-1} \sum_{j=1}^N (x_j^{(i)} - \mu_i)(x_j^{(i)} - \mu_i)^T, \quad i = w_1, w_2 \quad (4)$$

where $x_j^{(i)}$ is the j th sample vector from the i th class.

2.1.1. Class Conditional Mixture Density Estimation

Given the data sets for *lip* and *non-lip* classes from the previous frame, we can form the class conditional mixture density function in general as follows.

1. Form a 6-dimensional attribute data set for each class from color and texture measures (R, G, B, R_v, G_v, B_v) for each pixel location, and cluster it (possibly into three clusters for lip, tongue, and teeth) using an unsupervised K-means clustering algorithm.
2. Form the parametric class conditional density models $P(x | w_L^{(i)})$ using the method described in Section 2.1 for each cluster, where i represents the cluster i.d.
3. Similarly, repeat step 2-6 to form the parametric class conditional density models $P(x | W_{nL}^{(i)})$ for non-lips (nL).
4. Form the conditional density mixture models using weighted sum of the conditional densities belonging to clusters. That is,

$$P(x | w_i) = \sum_{m=1}^C c_m P(x | w_i^{(m)}), \quad i = L, nL \quad (5)$$

where C is the number of cluster for the lip or non-lip class, and $c_m = n_m/N$ is the mixture weight obtained by taking the ratio of the number of pixels in cluster m to total number of pixels in that class.

2.1.2. A Priori Probability Estimation

As shown in Equation 10, *a priori* probability specification is an important task for a Bayesian classifier since the *threshold value* of the likelihood ratio is based on the *a priori* class probabilities. Basically, it is desired to obtain a speaker and time (frame) dependent Bayesian parameter set to adapt the skin tone color variations and lighting variations on the fly. The selection of the sample data for obtaining class mean vectors and covariance matrixes has direct effect on the parametric representation of the class conditional density models. Calculating the *a priori* class probabilities based on the number of pixels in each class data is biased to the sample data so it would be a poor choice. By careful examination of the multi-variate Gaussian density function in Equation 1, one intuitional choice of the *a priori* class probabilities would be biasing them to determinant of the covariance matrixes of the classes, as

$$p(w_i) = \frac{\|\Sigma_i\|}{\|\Sigma_1\| + \|\Sigma_2\|}, \quad i = w_1, w_2 \quad (6)$$

where $p(w_1) + p(w_2) = 1$. Figure 1 shows the class regions based on the *threshold value* of the likelihood ratio (Bayes decision rule) and the effect of *a priori* class probability selection.

2.2. Bayesian Decision Rule

Let x be an observation vector (a set of features belong to a pixel location in the image frame). Our goal is to design a Bayes classifier to determine whether x belongs to w_1 or w_2 . The Bayes test using *a posteriori* probabilities may be written as follows:

$$p(w_1 | x) \underset{w_1}{\overset{w_2}{\gtrless}} p(w_2 | x), \quad (7)$$

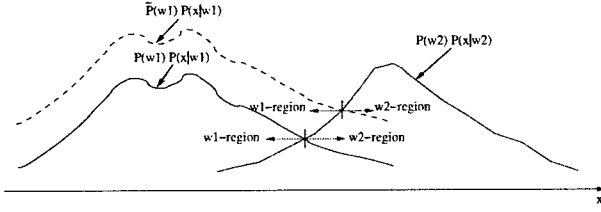


Figure 1: Bayes decision rule and the effect of the *a priori* class probability values.

where $p(w_i | x)$ is a *posteriori* probability of w_i given x . Equation 7 shows that, if the probability of w_1 given x is larger than the probability of w_2 , then x is declared belonging to w_1 , and vice versa. Since direct calculation of $p(w_i | x)$ is not practical, we can re-write the *a posteriori* probability of w_i using the Bayes theorem in terms of a *a priori* probability and the conditional density function $p(x | w_i)$, as

$$p(w_i | x) = \frac{p(x | w_i)p(w_i)}{p(x)} \quad (8)$$

where $p(x)$ is the mixture density function, and is positive and constant for all classes. Then, the decision rule shown in Equation 7 can be written as

$$p(x | w_1)p(w_1) \stackrel{w_2}{\lessgtr} p(x | w_2)p(w_2). \quad (9)$$

or re-arranging both sides, we get

$$L(x) = \frac{p(x | w_1)}{p(x | w_2)} \stackrel{w_2}{\lessgtr} \frac{p(w_2)}{p(w_1)} \quad (10)$$

where $L(x)$ is called the *likelihood ratio*, and $p(w_2)/p(w_1)$ is called the *threshold value* of the likelihood ratio for the decision. As shown in Equation 10 *a priori* probability specification is an important task for a Bayesian classifier. Because of the exponential form of the involved densities in Equation 10, it is preferable to work with the monotonic functions called discriminant functions following discriminant functions obtained by taking the logarithm of both sides of the Equation shown in 9.

$$q_i(x) = \ln(p(x | w_i)p(w_i)), \text{ or} \quad (11)$$

$$q_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) + \ln p(w_i) + c_i \quad (12)$$

where $c_i = -(1/2) \ln 2\pi - (1/2) \|\Sigma_i\|$ is a constant. In general Equation 12 has a nonlinear quadratic form and using Equation 12, the Bayes rule is as follows, which is preferable for the efficiency of calculation speed.

$$q_1(x) \stackrel{w_2}{\lessgtr} q_2(x). \quad (13)$$

2.3. Lip Tracking Algorithm and ROI Selection

The Bayesian framework described in this paper utilizes color images with no prior labeling. The goal is to segment the lip region in the current frame and select the ROI for the following frame to limit the search space. The basic lip tracking and ROI selection procedures are described below.

- Obtain $q_1(x)$ and $q_2(x)$ using Equation 11 for every pixel in the image.
- Use an averaging filter on the $q_1(x)$ and $q_2(x)$ to obtain $\{S_1(x)\}$ and $\{S_2(x)\}$. The smoothing operation reduces the noise effect.
- Apply the Bayesian classification rule to every pixel in the image frame to obtain binary lip candidate pixels, as

$$S_1(x) \stackrel{w_2}{\lessgtr} S_2(x). \quad (14)$$

- Segment the lip region (using the heuristics such as largest region between nostrils and chin) in the binary image resulted from the Bayes classifier.

The Bayesian classifier is applied to the full image array for the first frame. But once the lip region is detected on the current frame, the next frame's search space is bounded by a rectangular ROI, obtained by enlarging the current lip region by 25% of width and height in vertical and horizontal directions, respectively. Thus, the Bayesian classifier is applied to the ROI on the next frame to enable the real time lip tracking instead of the full image array search.

Adapting classifier parameters on the fly makes algorithm more robust to lighting changes between frames. Also the initial color information extracted from the first image frame may have several problems with changing conditions. Firstly, the color features obtained for a person by a camera is influenced by the ambient lighting conditions and orientation of the speaker's face during speech. Secondly, different cameras produce significantly different color features even for the same person under same lighting conditions. Our work aims to overcome this difficulty by adapting the classifier parameters on the fly using the information from the previous frame. The procedure is described as

- Extract the color features for *lip* class.
- Extract the color features for *non-lip* class.
- Update the classifier parameters using the data obtained from above two steps.

3. Removing Affine Parameters from Lip image

In the audio-visual speech and speaker recognition task, both contour based and pixel based visual features need to be independent from the affine (rotation, scale, shear and translation) parameters. In order to utilize the audio-visual speech and speaker recognizer for natural speech, the lip image for every frame needs to be pre-processed for removing the affine parameters before the visual feature extraction process described in the following sections is applied. Then, a question can be posed whether if affine (rotation, scale, shear and translation) parameters convey linguistic information to utilize for the recognition task.

3.1. Lip-Rotation Problem

Lip-rotation correction on the fly for natural speaker movement is essential for robust audio-visual speech and speaker recognition. Utilizing lip corners or some other facial features such as nostrils and eye corners may be problematic for rotation correction due to the complexity of locating such facial features accurately during natural speech [9, 12].

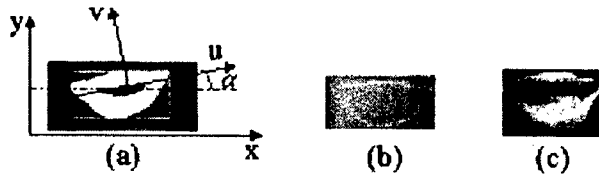


Figure 2: Lip rotation correction: a) rotation correction using the PCA, b) outer lip contour after rotation correction, c) gray lip image after rotation correction and scaling to 96x64 pixels.

We propose a principal component analysis (PCA) based rotation estimation and correction method to overcome the difficulties mentioned above. Jump

3.1.1. Rotation Correction Using PCA

Principal component analysis (PCA) is a method for analyzing multivariate data to identify a set of new orthogonal axes known as principal components. The first principal component is the axis that describes most variance of the data, the second principal component is the orthogonal axis that describes the second most variance of the data, and so on. PCA is also called the Hotelling transform or Karhunen-Loève expansion [13].

Let $\mathbf{x} = [x_1 x_2]^T$ be a 2-dimensional random variable with mean m_x and covariance matrix C based on N samples of a lip image pixel locations. The mathematical representation of PCA as follows.

$$m_{xk} = \frac{1}{N} \sum_{i=1}^N x_{ki}, \quad k = 1, 2 \text{ so} \quad (15)$$

$$m_x = [m_{x1} \ m_{x2}]^T \quad \text{and} \quad (16)$$

$$C = \frac{1}{N-1} \sum_{i=1}^N (x_i - m_x)(x_i - m_x)^T, \quad (17)$$

where T represents the transpose operation. The task is to find the new set of orthogonal axes and estimate the rotation angle with the standard coordinate system, and then undo the rotation of the lip pixel coordinate data. Figure 2 shows the rotation correction using the PCA coordinate rotation.

In order to estimate the rotation angle α between x -axis and u -axis shown in Figure 2a, we solve for the eigenvalues $\{\lambda_1, \lambda_2\}$ of the covariance matrix C and find the eigenvector e_1 corresponding to the largest eigenvalue. The process is as follows:

$$|C - \lambda I| = 0, \quad (18)$$

and then find the eigenvectors (also called proper vector or characteristic vector), calculated as

$$C e_i = \lambda_i e_i, \quad i = 1, 2 \quad (19)$$

where $e_1 = [e_{x1} \ e_{y1}]^T$. The eigenvector belongs to largest eigenvalue defines the rotation angle α , as

$$\alpha = \text{atan}(e_{y1}/e_{x1}). \quad (20)$$

Then the rotation correction matrix R^{-1} can be written as

$$R^{-1} = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}. \quad (21)$$



Figure 3: An example of the scaling problem due to speaker's distance to camera or speaker's lip physical dimensions.

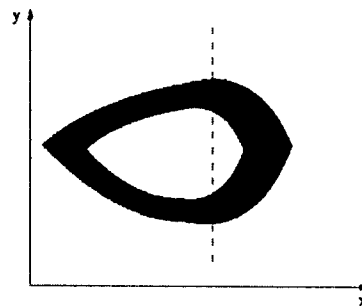


Figure 4: An illustration of the shearing in the horizontal direction.

The rotation corrected lip image is obtained by multiplying R^{-1} with the coordinates of lip pixel locations, as

$$\begin{bmatrix} x'_n \\ y'_n \end{bmatrix} = R^{-1} \begin{bmatrix} x_n \\ y_n \end{bmatrix}, \quad n = 1, 2, \dots, N \quad (22)$$

where $(x_n, y_n)^T$ represents the cartesian coordinates of the lip pixel locations, and $(x'_n, y'_n)^T$ represents the cartesian coordinate of the lip pixel locations after the rotation correction. Figure 2c shows the orientation of the lip shape after rotation correction and scaling of lip shown in Figure 2a.

3.2. Scaling Problem

The scaling problem occurs due to the speaker's distance to camera, the camera zoom factor and the speaker's actual lip dimensions. In this case, any pixel based visual feature extraction method such as DCT or wavelet transform method which utilizes the frequency content of the lip image may generate inconsistent (noisy) observation vectors. To overcome this problem, we propose to interpolate every lip image to same size, $N \times M$. Figure 3 shows the scaling problem example for two different speakers and the lip images of them after interpolation (scale correction).

3.3. Shearing (Uneven Scaling) Problem

Shearing occurs when the speaker's head position is not perpendicular to camera optical axis. For example, one side of the lips which may look larger than the other. Solving the shearing problem using the single 2D image information is not theoretically possible. There can be various practical approaches to minimize the shearing effect such as using the symmetry information of the lips may enable us to estimate the shear matrix by utilizing the least squares estimate method and undo the shearing. Figure 4 illustrates a typical example of a shearing effect in the horizontal direction.

The shearing may also be associated with the accent of a speaker, depending on certain visemes. Then, the similar

question can be posed whether shearing conveys a linguistic information.

4. Visual Speech Feature Extraction

Lipreading clearly meets at least two practicable criteria: It mimics human visual perception of speech recognition, and it contains information that is not always present in the acoustic signal [3, 4, 14–16]. Petajan is one of the first researchers who built a lipreading system using oral-cavity features to improve the performance of an acoustic ASR system [17]. Silsbee et al. [18] utilized vector quantization (VQ) of acoustic and visual data for their HMM based audio and video subsystems. Teissier et al. [19] utilized 20 FFT based 1-bark wide channels between 0 and 5 KHz for acoustic features and inner lip horizontal width, inner lip vertical height and inner lip area for the visual features. Chiou et al. [20] utilized active contour modeling to extract visual features of geometric space, the Karhunen-Loève transform (KLT) to extract principal components in the color eigenspace, and HMMs to recognize the combined video only feature sequences. Potamianos et al. [14, 21] used Fourier descriptor magnitudes for a number of Fourier coefficients, width, height, area, central moments, normalized moments as contour features, image transform features, and hierarchical discriminant features.

In order to utilize audio-visual ASR for natural speech in varying lighting conditions, the visual front end algorithm that extracts the visual features must satisfy the three criteria presented in Section 1. The contour based feature described in Section 4.1 satisfy step (i) in the Fourier domain and is relatively independent of step (ii) and step (iii). For pixel based visual feature extraction methods, step (i) is explained in Section 3. Steps (ii) and (iii) are explained for both 2D DCT based visual features and kurtosis measure based visual features which are described in Sections 4.2, and 4.3, respectively.

4.1. AI-FDs Based Visual Features

In general, for the video feature extraction, the relationship between observed parametric outer-lip contour data \mathbf{x} and parametric reference data \mathbf{x}^o can be written as,

$$\mathbf{x}[n] = A\mathbf{x}^o[n + \tau] + \mathbf{b}, \quad (23)$$

where A represents a 2×2 arbitrary affine matrix, $\det(A) \neq 0$, that may have scaling, rotation, and shearing affect, \mathbf{b} represents a 2×1 arbitrary translation vector, and τ is starting point. These are removed in the Fourier domain [7, 22]

The video feature extraction algorithm extracts twelve affine-invariant Fourier descriptors (AI-FDs) of the parametric outer lip contour data as well as four affine-invariant oral cavity features which are width, height, ratio of width to height, and outer lip's inner area by normalizing the next frame's corresponding oral cavity features. Dynamic coefficients, which are used as a video observation features, are obtained by differencing the consecutive image sequence features.

4.2. Normalized 2D DCT Based Visual Features

The Discrete Cosine Transform is one of the many transform methods that transforms its input into a linear combination of weighted basis functions. The 2D DCT on a $N \times N$

lip image can be written as

$$Y = C^T X C \quad (24)$$

where X is an $N \times N$ lip image, Y contains the $N \times N$ DCT coefficients, and C is an $N \times N$ transform matrix defined as

$$C_{mn} = k_n \cos\left[\frac{(2m+1)n\pi}{2N}\right], \text{ where} \quad (25)$$

$$k_n = \begin{cases} \sqrt{1/N} & \text{when } n = 0, \\ \sqrt{2/N} & \text{otherwise} \end{cases}$$

and $m, n = 0, 1, \dots, N-1$. Our goal is to extract visual features satisfying step (ii) and step (iii), and most relevant information of the lip shape from the $N \times N$ DCT coefficients. Let I^o and I be lip shape images which differ in a scale and shift factors (lighting condition). i.e.,

$$I = \alpha I^o + \delta, \quad (26)$$

where α and δ are scale and shift factors in the acceptable range¹ of the chrominance/luminance space.

From Equation 25, we know that the zeroth coefficient of the DCT transform contains the DC information (δ in Equation 26) which doesn't convey any shape information. It is also known that DCT is a linear transform and the scale factor α just scales all the DCT coefficients. So normalizing all the coefficients in the DCT domain by a coefficient Y_{mn} makes the DCT transform scale independent. Then, 35 coefficients from the lower frequencies are selected excluding the DC information. Figure 5 shows the normalized 2D DCT based visual feature extraction process.

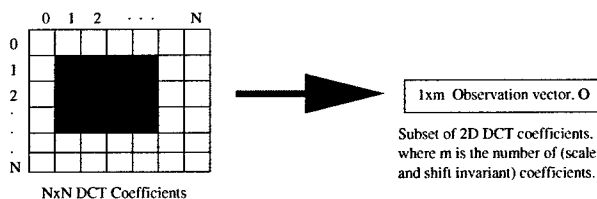


Figure 5: Normalized 2D DCT based visual feature extraction.

4.3. 2D Kurtosis Measure of the Probability Density Distribution of the DCT Coefficients

After the rotation correction and size normalization of the lip image, the resulting lip image is divided into 16×16 sub-blocks with 50% overlapping or non-overlapping sub-blocks, and then the two-dimensional DCT of the each block is calculated. For simplicity, let Y be the matrix of 16×16 DCT coefficients. $Y(0,0)$ depends only on the chrominance/luminance space *shift* shown in Equation 26, and conveys no shape information. Thus, the $Y(0,0)$ coefficient is removed. The remaining coefficients are now only chrominance *space* scale dependent (see Equation 26). We remove the dependency on the chrominance space *scale* by calculating the 2D kurtosis of the frequency profile (probability distribution of DCT coefficients) of each block in the lip image discussed in the following sections. Figure 6 shows the pixel

¹Reference and observed lip image contents are clearly visible for a range of α and δ .

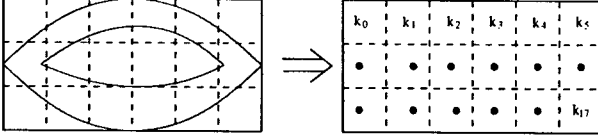


Figure 6: Illustration of FPM visual feature extraction (k_i is an appearance based visual coefficient for the i th lip image block).

based visual front end process, where k_0, k_1, \dots, k_{17} are coefficients for the pixel (appearance) based visual features of the lip image. In this work, we will refer these pixel based features as frequency profile measures (FPMs), which are 2D kurtosis measures of the probability density distribution of the DCT coefficients.

In the theory of probability, the classical measure of the non-Gaussianity of a random variable is the kurtosis measure. Kurtosis measures the departure of a probability distribution from the Gaussian (normal) shape². Kurtosis is dimensionless ratio, and greater than zero for most non-Gaussian random variables³. Specifically, for a given 2D image block function $I(n, m)$, where $m, n = 0, 1, \dots, N$, the corresponding 2D DCT coefficients $Y(x, y)$ can be obtained as described in Section 4.2, where x and y are the spatial frequencies in the DCT domain. The high-frequency DCT coefficients⁴ are discarded to minimize the video noise effect which is discussed in Section 4.3.1. The rest of the lower frequency DCT coefficients $Y(x, y)$ for $x, y = 1, 2, \dots, N/2$, are normalized to form the bi-variate probability density function $p(x, y)$. Using the notation of [23], for a given univariate random variable x with marginal probability mass function $p(x)$, mean μ_x , and existing finite moments up to the fourth moment, then, the univariate kurtosis is defined by:

$$\text{kurt}(x) = \beta_2 = \frac{m_4}{m_2^2}, \quad (27)$$

where m_2 and m_4 are the second and fourth central moments, respectively. In general, the k th central moment is defined by:

$$m_k = E[(x - \mu_x)^k] = \sum_x (x - \mu_x)^k p(x), \quad (28)$$

where marginal density function of x is

$$p(x) = \sum_y p(x, y), \quad (29)$$

where E denotes the probability expectation [24]. If x_1 and x_2 are two independent random variables, then kurtosis has the following linearity properties:

$$\text{kurt}(x_1 + x_2) = \text{kurt}(x_1) + \text{kurt}(x_2) \quad \text{and} \quad (30)$$

$$\text{kurt}(\alpha x_1) = \alpha^4 \text{kurt}(x_1) \quad (31)$$

where α is an arbitrary scalar. Clearly, any scale factor in Equation 27 cancels out. Let W be a p -dimensional random vector with finite moments up to the fourth, and μ and

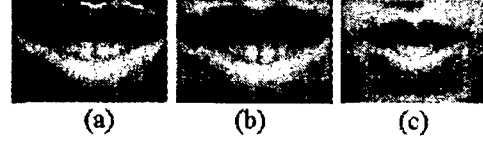


Figure 7: In search of the lip region type with 96x64 pixel size to extract visual speech features: a) exact lip region, b) exact rectangular lip region, c) extended rectangular lip region.

Γ be the mean vector and covariance matrix of W , respectively. Mardia [25] proposed the p -dimensional multivariate kurtosis as:

$$\beta_{2,p} = E[(W - \mu)^T \Gamma^{-1} (W - \mu)]^2, \quad (32)$$

where T denotes the transpose of a vector. Zhang [23] used 2D kurtosis of random vectors for a sharpness measure of Scanning Electron Microscopy (SEM) images. The 2D kurtosis $\beta_{2,2}$ is calculated by

$$\beta_{2,2} = [\gamma_{4,0} + \gamma_{0,4} + 2\gamma_{2,2} + 4\rho(\gamma_{2,2} - \gamma_{1,3} - \gamma_{3,1})] / (1 - \rho^2)^2, \quad (33)$$

where

$$\gamma_{k,l} = \sum_x \sum_y (x - \mu_x)^k (y - \mu_y)^l p(x, y) / [(\sum_x (x - \mu_x)^2 p(x))^{k/2} (\sum_y (y - \mu_y)^2 p(y))^{l/2}], \quad (34)$$

$$\sigma_{xy}^2 = E[(x - \mu_x)(y - \mu_y)], \quad \sigma_x^2 = E[(x - \mu_x)^2], \quad (35)$$

and

$$\rho = \sigma_{xy}^2 / (\sigma_x \sigma_y). \quad (36)$$

The 2D kurtosis measure, $\beta_{2,2}$, is dimensionless and *scale* and *shift* invariant as seen in Equation 33. In this work, the 2D kurtosis defined in Equation 33 is calculated using the probability density distribution of the DCT coefficients of the image block function $I(n, m)$. We will refer to the $\beta_{2,2}$ measure as the frequency profile measure (FPM) of an image block. The image blocks, which have zero marginal variances of x or y , are discarded for $\beta_{2,2}$ calculation, and their FPMs are arbitrarily assigned to the $\gamma_{4,0}$ value when $\sigma_x \neq 0$ and $\sigma_y = 0$, to the $\gamma_{0,4}$ value when $\sigma_y \neq 0$ and $\sigma_x = 0$, and to -1 when both $\sigma_y = 0$ and $\sigma_x = 0$.

4.3.1. Reducing the Effect of Video Noise in FPM Visual Features

It is known that the low-frequency coefficients in the DCT of the video signal contain the large details and the high-frequency coefficients contain the finer details of the image. Video noise⁵ is clearly represented in the DCT coefficients and using the full spectrum of the image leads to noisy (distorted) visual features. That is why some of the high-frequency DCT coefficients were discarded in the calculation of FPM of the image blocks described in Section 4.3. The pixel based visual front end research requires further investigation on how to minimize the effects of video noise and the dependence of FPM on the selection of the cut-off frequency.

²The smaller the kurtosis, the flatter the top of the distribution.

³Kurtosis is 3 for any univariate Gaussian distribution.

⁴The upper half of the DCT coefficients are discarded.

⁵Motion blur, coding artifacts, quantization errors, electronic noise, etc., are considered to be video noises.

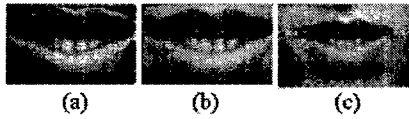


Figure 8: In search of the lip region type with 80x48 pixel size to extract visual speech features: a) exact lip region, b) exact rectangular lip region, c) extended rectangular lip region.

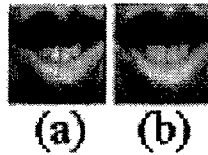


Figure 9: Effect of interpolating on pixel based visual feature extraction: a) re-interpolated from 96x64 pixels to 60x60 pixels, b) re-interpolated from 80x48 pixels to 60x60 pixels.

Table 1: Visual-only recognition accuracy for connected digit task using the subset of the normalized 2D DCT features, FPM features, and concatenated AI-FDs and FPM features. (LR: lip region, R-LR: rectangular LR, ER-LR: extended R-LR, bl.: blocks).

Sub. of norm. 2D DCT using	TR V%	TS V%
exact LR with ini. 80x48 pixels	22.40	21.60
exact LR with ini. 96x64 pixels	23.00	20.80
R-LR with ini. 80x48 pixels	24.60	17.20
R-LR with ini. 96x64 pixels	24.00	19.60
ER-LR with ini. 80x48 pixels	22.80	24.40
ER-LR with ini. 96x64 pixels	21.60	21.60
FPMs using		
exact LR with overlapping bl.	41.80	19.60
exact LR with non-overlapping bl.	35.00	24.00
R-LR with overlapping bl.	38.80	23.60
R-LR with non-overlapping bl.	34.60	22.00
ER-LR with overlapping bl.	39.00	22.00
ER-LR with non-overlapping bl.	34.20	19.60
Concat. AI-FDs and FPMs using		
only AI-FDs	18.55	21.33
exact LR with overlapping bl.	19.20	18.40
exact LR with non-overlapping bl.	17.60	18.40
R-LR with overlapping bl.	18.40	20.40
R-LR with non-overlapping bl.	17.40	18.40
ER-LR with overlapping bl.	18.40	17.60
ER-LR with non-overlapping bl.	17.80	18.80

5. Visual-Only Experimental Setup and Results

This paper discusses visual modality speech recognition (lipreading) system setup and results. The HMM states were modeled with continuous density Gaussians with single mixture components. The aim of this work is to investigate an affine and lighting conditions invariant visual feature extraction method. Therefore, the HMM model structure was kept basic. The HMM implementation was word level, left-to-right with no skip transitions with ten (eight emitting and two non-emitting) states, and diagonal covariance Gaussian mixture components since we assume that the coefficients in the observation vectors are naturally independent. All the model parameters were initialized using the Viterbi training algorithm and re-estimated using the Baum-Welch re-estimation algorithm. Viterbi recognition (dynamic programming) algorithm is utilized for the recognition.

The Clemson University Audio-visual Experimental (CUAVE) connected and continuous audio-visual digit database, which is a thirty six subject dataset, was utilized for the experiment. The visual-only experimental results are presented for a connected audio-visual digit recognition task. The following visual features from exact lip region, exact rectangular lip region, and generous rectangular lip region as shown in Figures 8 and 9 are utilized in the visual-only speech recognition system.

1. Subset of normalized 2D DCT features
2. FPM features
3. AI-FD features
4. Concatenated AI-FDs and FPM features

The subset of the 36 speaker dataset, containing 15 speakers each is uttering five times 0-9. The speakers are split into training (TR) and testing (TS) set of ten and five subjects, respectively, leading to speaker independent visual only recognition system. The results are shown in Table 1.

6. Concluding Remarks and Future Work

Table 1 shows the visual-only connected digit recognition results, where TR corresponds to training set performance and TS corresponds to test set performance, for various visual features discussed in this paper. The subset of the normalized 2D DCT features based on the training set results from exact rectangular lip region gives better results than the exact lip region and extended lip region (see in Figure 9). Another observation is that slight change in lip image content due to the linear interpolation has effects on the system's performance.

In the results obtained using FPM features, the training set performance is much better than the test set performance. Similar performance behavior was observed for a speaker dependent recognition task. Therefore, we conclude that FPM based features are highly video noise sensitive. The overlapping block based FPM features outperformed the non-overlapping block based FPM features significantly in the training set. Among the three different lip regions shown in Figure 9, the exact lip region with overlapping blocks method outperforms the results of outer two regions.

In the results obtained using concatenated AI-FDs and FPMs. the training set and test set performances are close

to each other and worse than FPMs-only results. Therefore, we conclude that each feature should be treated as a separate stream and weighted properly to bring the additional information from one another. Similarly, the slight performance increase due to the overlapping block of FPM features over non-overlapping block based FPM features can be noticeable.

We also report that the number of mixtures in the Gaussian mixture model (GMM) selection and the number of states in the silence model affects the performance of visual-only system. For example, setting GMM to twelve and using embedded training of the FPM based visual only system achieved 98% recognition accuracy on the training set, but about 16% on the speaker independent test set (which is less than the result of single GMM reported in Table 1. The similar behavior is observed for the speaker dependent set. That is, the system is being well trained with the FPM features, but the both test sets are behaving like an unmatched system due to the resulting noisy observations.

We conclude that visual noise is an important factor in visual speech feature extraction, and overlapping local image block based FPM features outperform normalized 2D DCT features, AI-FD features, and concatenated AI-FDs and FPM features. Future work will include initial lip segmentation for the Bayesian framework training and further study on the noise robust FPM feature extraction.

7. References

- [1] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition," in *Proceedings of ICASSP*, 2002.
- [2] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Speech Reading by Humans and Machines*, D. G. Stork and M. E. Hennecke Eds. Springer, Berlin, 1996.
- [3] E. Vatikiotis-Bateson, K. G. Munhall, M. Hirayama, Y. V. Lee, and D. Terzopoulos, "The dynamics of audio-visual behavior in speech," in *Speechreading by Man and Machine: Data, Models and Systems*, D. G. Stork and M. E. Hennecke Eds. NATO Springer-Verlag, New York, NY (1996), 1996, vol. 150.
- [4] S. Nakamura, "Fusion of audio-visual information for integrated speech recognition," in *Audio- and Video-Based Biometric Person Authentication*, 2001.
- [5] C. Neti, G. Potamianos, J. Luetttin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," in *JHU Workshop 2000*. <http://www.clsp.jhu.edu/ws2000/>, 2000.
- [6] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration," *American Scientist*, vol. 86, 1998.
- [7] S. Gurbuz, E. Patterson, Z. Tufekci, and J. Gowdy, "Lip-reading from parametric lip contours for audio-visual speech recognition," in *Proceedings of Euro Speech*, 2001.
- [8] A. W. Senior, "Face and feature finding for face recognition system," in *Proceedings of AVBPA*, 1999, pp. 154-159.
- [9] G. Iyengar, G. Potamianos, C. Neti, T. Faruque, and A. Verma, "Robust detection of visual roi for automatic speechreading," in *IEEE Fourth Workshop on Multimedia Signal Processing*, 2001, pp. 79-84.
- [10] A. Yuille, "Feature extraction from faces using deformable templates," *Int. Journal of Computer Vision*, 8(2), pp. 99-111, 1992.
- [11] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," in *Int. Proc. 1st Int. Conf. on Computer Vision*, 1987, pp. 259-268.
- [12] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "Moving-talker, speaker-independent feature study and baseline results using the cuave multimodal speech corpus," *EURASIP Journal on Applied Signal Processing* (accepted for publication), 2002.
- [13] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, 1997.
- [14] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for hmm-based automatic lipreading," in *Proceedings of ICIP*, 1998.
- [15] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Proceedings of ICASSP*, 1993.
- [16] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an hmm-based asr," in *Speechreading by Humans and Machines: models, systems, and applications*, NATO ASI Series. Series F, Computer and Systems Sciences no. 150, pp. 461-471, 1996.
- [17] E. Petajan, B. Bischoff, D. Bodoff, and N. Brooke, "An improved automatic lipreading system to enhance speech recognition," in *In CHI 88*, pp. 19-25, 1988.
- [18] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Transactions on Speech, and Audio Processing*, vol. 4, no. 5, 1996.
- [19] P. Teissier, J. Robert-Ribes, J. Schwartz, and A. Guérin-Dugué, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Transactions on Speech, and Audio Processing*, vol. 7, no. 6, 1999.
- [20] G. I. Chiou and J. Hwang, "Lipreading from color video," *IEEE Transactions on Image Processing*, vol. 6, no. 8, 1997.
- [21] G. Potamianos, J. Luetttin, and C. Neti, "Hierarchical discriminant features for audio-visual lvcsr," in *Proceedings of ICASSP*, 2001.
- [22] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition," in *Proceedings of ICASSP*, 2001, vol. 1, pp. 177-180.
- [23] N. F. Zhang, M. T. Postek, R. D. Larrabee, A. E. Vladar, W. J. Keery, and S. N. Jones, "Image sharpness measurement in scanning electron microscope - part iii," in *Scanning*, 1999, vol. 21, pp. 246-252.
- [24] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, Inc., 1991.
- [25] K.V. Mardia, "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, vol. 57, pp. 519-530, 1970.

Performance Analysis of Automatic Lip Reading Based on Inter-Frame Filtering

Jinyoung Kim*, Seungho Choi**, Seongmo Park*

*Dept. of Electronics, Computer and Information Engineering, Chonnam National University, Kwang-ju, Korea

**Dept. of Information & Communications Engineering, Dongshin University, Naju, Korea

kimjin@dsp.chonnam.ac.kr, smpark@chonnam.ac.kr, shchoi@white.dongshinu.ac.kr

Abstract

Automatic lip-reading has been focused as a complimentary method of automatic speech recognition in noisy environments. One of the most competitive lip-reading algorithms is the image transform based lip-reading (ITLR) algorithm. However, ITLR has severe performance degradation under illumination variations.

RASTA is a kind of inter-frame filtering method. It is used for rejecting stationary and convolutional noise in speech signal processing. In this paper, we apply RASTA approach to ITLR and analyze the performance of this method. We propose two merging techniques of pre-integration (PRE-I) and post-integration (POST-I). In PRE-I RASTA, inter-frame filtering is performed ahead of the image transform process. In POST-I, inter-frame filtering is done after the image transform process. We also compare the effectiveness of high-pass filtering and band-pass filtering as inter-frame filtering.

Experimental results show that pre-integration is very effective to reject illumination variances. And it is observed that high-pass filtering is enough to enhance the performance of lip-reading.

1. Introduction

Recently, researches on automatic lip-reading using the video sequence of the speaker's mouth have attracted significant interest. Automatic lip-reading under noisy environments is very effective in compensation for the decrease of speech recognition rate with an audio-only speech recognition (ASR) system [1]. The bimodal based on audio-visual information is an important part of the human-computer interface (HCI). We allow more weighting value to visual data than to audio one under a bad SNR but, on the contrary, more to audio data than to visual one under a clean SNR [2]. Under noisy circumstances, this bimodal approach has been a good alternative showing superior recognition rate to audio-only ASR system.

In this paper, we concentrate on the image transform based approach for automatic lip-reading (ALR) for bimodal speech recognition system. This approach is known to be superior to a lip-contour-based method for visual-only HMM recognition tasks. However, while the lip-contour based approach needs only several visual data, for example, outer, inner lip contour and lip width, the image-transform-based approach requires much larger visual feature vectors since it is based on the whole transformed image data of the speaker's mouth. Thus, for a fast algorithm, the necessity to reduce those data size has arisen.

To reduce the dimensionality of feature vectors, principal components analysis (PCA) has been suggested as a good method,

which is based on linearly projecting the image space to a low dimensional feature space [3]. By the way, ITLR has the problem of robustness. Under varying illumination, the observed image sequences are suffered from rapid performance degradation. Illumination variation from the inconsistency of training and test conditions interferes the recognition process such as exact feature extraction. This interference causes a mismatching between the correct word and the related feature model and, after all, reduces the recognition rate. Our preliminary experiment in lip-reading system showed that even only a small amount of intensity variation caused large degradation of lip-reading performance [4].

To tackle those problems we propose the inter-frame filtering method, which is very similar with RASTA filtering in automatic speech recognition (ASR). According to reference [5], RASTA filtering is very successful in ASR under convolutional noisy environment. We propose two kinds of integration methods, pre-integration and post-integration. We examine usefulness of the inter-frame approach with our own lip-reading system.

In section 2, we briefly describe the algorithm for real-time automatic visual-only lip-reading system and mention about the necessity of the proposed method. Section 3 describes methods to diminish the illumination noise for the improved recognition rate. Finally, section 4 presents experimental results.

2. Baseline system : visual-only HMM-based lip-reading system

To develop a robust lip-reading algorithm, we implemented an automatic image transform based lip-reading system using HMM based word model. Figure 1 shows the overall block diagram of the implemented system based on the proposed algorithm. Given image sequence containing speaker's mouth, the overall process to extract the visual feature data consists of two sub-processes. One is ROI (region of interest) extraction process and the other is feature parameter extraction process.

2.1 ROI extraction

Since lip-reading is based on the visual information of moving lip, extraction of appropriate interesting regions containing only moving lip area is important. ROI extraction from each image frame of given sequence is performed before feature extraction. As shown in figure 1, ROI extraction process consists of three steps; 1) gray-level transformation, 2) masking filtering and 3) binary-level transformation.

To find lip area efficiently, color image is first transformed into gray level image and then into binary-level image.

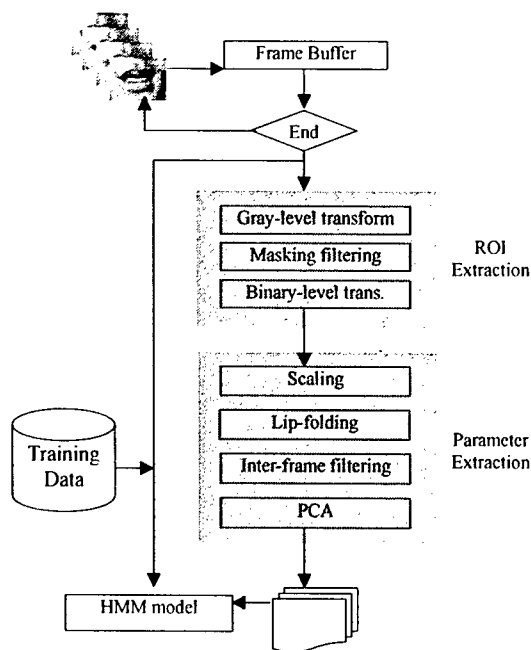


Figure 1. Block diagram of the proposed method for real-time visual-only HMM based lip-reading system

Both lip-ends of moving lip are extracted from this binary-level image by applying Y-projection and then X-projection. The vertical and horizontal center of speaker's mouth is obtained from these X, Y-projection. Then, the square pixel window of ROI is constructed around speaker's mouth. Since the lip width information of moving lip is important, we keep the width of ROI obtained at the first frame of each word to the last frame of that word. During the ROI extraction process, 'masking filter' is applied to diminish the unbalanced illumination of facial area from various lighting source.

2.2 Feature extractions

To reduce the visual feature parameter size, each ROI is downsampled into a 16 x 16 pixel window for fast algorithm. This operation is necessary not only to reduce the feature data size but also to normalize the difference between each ROI size due to variations such as speaker's lip widths and the distances from camera.

To reduce the parameter size, dimensionality of visual feature vector, PCA (principal component analysis) is applied. PCA is known as a simple method to implement and to guarantee good performance in automatic lip-reading [6]. And, we use lip-folding technique before PCA process. Lip-folding is based on the symmetric property of lip along the vertical axis. Lip-folding makes 16 x 16 image size to half size of 8 x 16. The mean half-sized image needs smaller principal components to represent it than the original unfolded one. Additionally, the mean image compensates the illumination unbalance between the left lip area and the right lip area and, therefore, shows robustness under various lighting conditions[7].

2.3 HMM based word recognition

For every video field, a static observation feature vector is acquired and those vectors obtained from the given video

sequence are used for HMM based word modeling. Our automatic lip-reading system uses continuous density HMMs as a means of statistical pattern matching. The HMM observation probabilities are modeled as multi-dimensional Gaussian mixtures with diagonal covariance matrices. For the specific lip-reading recognition tasks considered in this paper, we use whole word, 3-6 state, left-to-right models with 3-8 mixtures per state. All HMM parameters are estimated by maximum likelihood Viterbi training.

3. Inter-frame filtering

One of ASR problems is the robustness. The performance of ASR is commonly worse in noisy environments. In general, noise is classified into additional and convolutional noise. RASTA filtering is one of methods used in ASR for preventing the degradation of ASR performance. RASTA is the abbreviation of 'relative spectral smoothing'. It was found that filtering time trajectories could compensate greatly for the effect of the convolutional noise induced by communication channel[5]. RASTA filtering is performed with bandpass filter. In RASTA filtering slow varying components, corresponding to the frequency characteristics of communication channel, are suppressed. The low-pass filtering helps to smooth some of the fast frame-to-frame spectral change present. The commonly used bandpass filter is as follows.

$$H(z) = 0.1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \quad (1)$$

Based on these results, we discuss how inter-frame filtering is applied to lip-reading problems to enhance the performance of automatic lip-reading.

3.1 Integration of inter-frame filtering with lip-reading system

According to original work of Hermansky, RASTA filtering is applied to speech feature vector (SFV) sequence after obtaining SFVs. The RASTA filter is a kind of bandpass filter to reject slow and fast varying components. In our lip-reading system, feature extraction processing is PCA and the feature parameters are projection values of original image into most important axis. Thus, we can integrate inter-frame filtering after PCA in our lip-reading system, a simple imitation of ASR structure adopting RASTA filtering. We call this approach as post-integration (Post-I). Figure 2 shows the block diagram of Post-I method.

On the other hand, our AV database (DB) was recorded at various lighting conditions with illumination not regulated when visual DB was recorded. Thus, we may think that our AV DB was originally suffered from illumination noise. If the illumination noise was variant and dynamic, the result of PCA may include the influence of illumination noise. So, the m important axes would contain the components induced by illumination noise. This concept makes us change the order of PCA and inter-frame filtering. Figure 3 shows the second integration method of pre-integration (Pre-I).

3.2. Filters for inter-frame filtering

The band-pass filter used in ASR is shown in eq. (1). It is not impossible to use this filter for filtering image sequence. It's

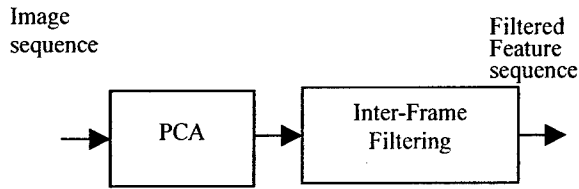


Figure 2. Post-integration method(Post-I).

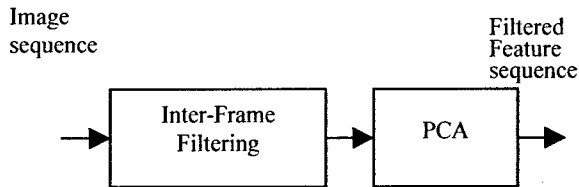


Figure 3. Pre-Integration method(Pre-I).

because the sampling frequency is very low in case of image capture operation compared with speech sampling. For speech signal 100 feature vectors per second is common. But, in our case, sampling frequency for image signal is 30Hz/second. So, we used very simple IIR filter for inter-frame filtering as follows.

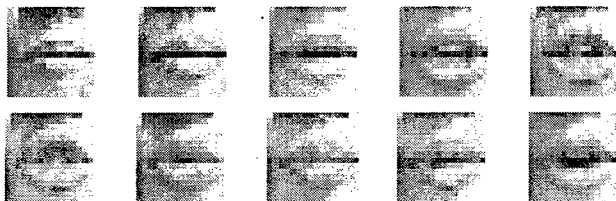
High-pass filter :

$$Y_i[n, m] = 0.9858 \cdot (X_i[n, m] - X_{i-1}[n, m]) + 0.9716 \cdot Y_{i-1}[n, m] \quad (2)$$

Low-pass filter :

$$Y_i[n, m] = 0.8638 \cdot (X_i[n, m] + X_{i-1}[n, m]) + 0.7257 \cdot Y_{i-1}[n, m] \quad (3)$$

Both filters are IIR(1,1) filters designed using MATLAB tool. Figure 4 shows the original image sequence and the filtered image sequences.



(a) Original image sequence (16 x 16)



(b) High-pass filtered image sequence (8 x 16)



(c) Band-pass filtered image sequence (8 x 16)

Figure 4. Inter-frame image filtering results

Table 1. Experimental environments.

Camara	SONY digital home video camera
Frame rate	30 frames/sec
Words	22 Korean words selected from the command menu for car navigation system
Training speakers	52 male speakers
Test speakers	18 male speakers different from training speakers
Recording condition	All recording are performed at different rooms at different time

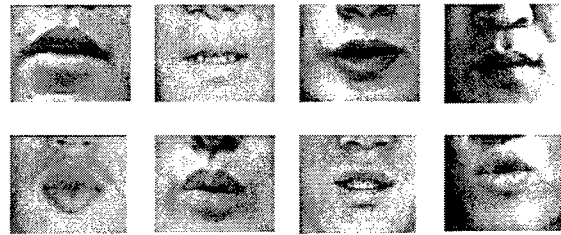


Figure 5. Some examples of our database recorded.

4. Experimental Environments and Results

4.1 Experimental environments

The experimental environment is shown in table 1. The database is composed of 22 Korean words spoken by 70 speakers. Figure 5 shows sample images of the AV database. As shown in the figure, our database recorded at different rooms and at different time, reveals illumination variations.

4.2 Experimental results

In this subsection, we describe the results of two proposed integration methods; Pre-I and Post-I, in the point of feature vector dimension and recognition results. Table 2 shows the dimension of features in Pre-I and Post-I integrations. From table 2, it is observed that post integration method is very effective in

Table 2. Comparison of feature dimensions in cases of Pre-I and Post-I

	Filter	PCA 90%	PCA 95%
Post-Integration	Bandpass	24	44
	Highpass	24	44
	NonFilter	24	44
Pre-Integration	Bandpass	6	14
	Highpass	6	14
	NonFilter	24	44

reduction of principal component numbers. The reason for this achievement could be that the pre-filtering rejects the influence of illumination noise before PCA process.

The other observation is that the low-pass filtering does not reduce the feature vector dimension. This result is not remarkable, for the sampling rate of image signal is much lower than that of speech signal. Anyway, using the post-integration, the feature vector dimension is reduced up to approximately 30%. The recognition results are shown in figure 6 and 7. From these two figures we can observe the following facts.

- 1) The post-integration doesn't improve the lip-reading performance. It makes the lip-reading performance worse. But the pre-integration enhance the recognition rate of the lip-reading system. This fact is the different point compared with the ASR.
- 2) The band-pass filtering, especially low-pass filtering is not decisive to increase the recognition rate. In other words, high-pass filtering is enough to the lip-reading system. As discussed above, it's because the sampling rate of video data is high when we consider the rate of lip movements in speaking.

It is obvious that pre-integration of inter-frame filtering is very effective in automatic lip reading. Pre-integration not only reduces the dimension of feature space but also improves the recognition rate of image-based lip-reading system.

5. Concluding Remarks

In general, lip-reading performance, especially image transform based lip-reading, is very sensitive to illumination variance. So, it is necessary to develop the robust version of lip reading to use automatic lip-reading in real service environments.

In this paper, we proposed inter-frame filtering approach as one of robust lip-reading methods and analyzed the performance of the proposed methods. From our experimental results we showed that pre-integration of inter-frame filtering enhanced lip-reading performances. The achievements are as follows.

- 1) Inter-frame filtering reduced feature vector dimension.
- 2) Inter-frame filtering improved the recognition rate of automatic lip reading.

In the future work, we will enlarge our AV database and study more robust methods so that automatic lip-reading can be used in real environments

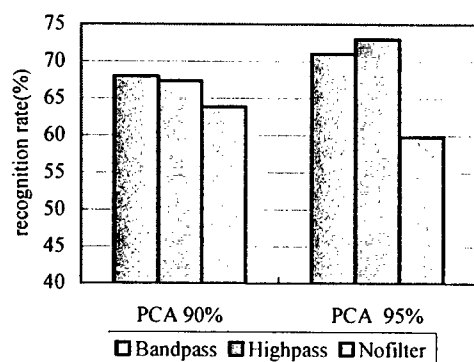


Figure 6. Recognition results of post-integration

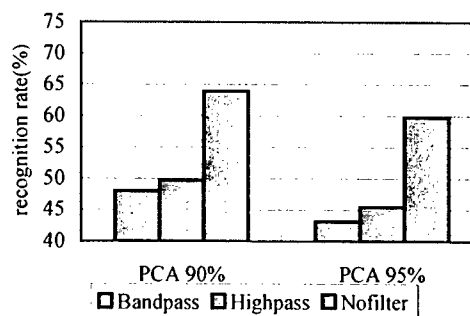


Figure 7. Recognition results of pre-integration

Acknowledgements

This work was supported partially by RRC HECS and ETRI.

References

- [1] Meier U., Hürst W., Duchnowski P., "Adaptive Bimodal Sensor Fusion For Automatic Speechreading", Proc. of ICASSP'96, Vol. 6. No. 2., pp. 833-836, 1996.
- [2] Sharma R., Pavlovic V. I., Huang T. S., "Toward Multimodal Human-Computer Interface", Proc. of IEEE, Vol. 86. No. 5., pp. 853-869, 1998.
- [3] Sirovitch L. and Kirby M., "Low-Dimensional Procedure for the Characterization of Human Faces", J. Optical Soc. of Am., Vol. 2, pp.519-524, 1987.
- [4] Kim J., Lee J. and Shirai K., "A Study on Various Factors Concerned with Lip-reading Performance at Dynamic Environment", J. of ICSP'01, pp.923-926, 2001.
- [5] Hermansky H., Morgan M. and Bayya A., and Kohn P., "Compensation for the effect of the communication channel and auditory analysis of speech (RASTA-PLP)," Proc. Of Eurospeech'91, pp.1367-1371, 1991.
- [6] Potamianos G., Graf H. P. and Cosatto E., "An Image Transform Approach for HMM Based Automatic Lipreading," Proc. of ICASSP'98, pp.173-177, 1998.
- [7] J. Lee and J. Kim, "An efficient Lipreading Method Using the Symmetry of Lip," Proceedings of EuroSpeech2001, pp.1019-1022, 2001.
- [8] Daubias P., Deleglise P., "Evaluation of an Automatically Obtained Shape and Appearance Model For Automatic Audio Visual Speech Recognition," Proc. of Eurospeech2001, pp.1031-1034, 2001.
- [9] Liévin M. and Luthon F. "Lip Features Automatic Extraction", Proc. Of the 5th IEEE Int. Conf. On Image Processing. Chicago. Illinois, 1998.
- [10] Uwe Meier, Rainer Stiefelbogen, Jie Yang, "Preprocessing of visual speech under real world conditions", Interactive Systems Lab. European Tutorial & Research Workshop on Audio-Visual Speech Processing: Computational & Cognitive Science Approaches (AVSP 97).

Robust Head Tracking Based on Hybrid Color Histogram and Random Walk Kalman Filter

Gwang-Myung Kim¹, DongCheng Lin², Jung H. Kim¹, Sung H. Yoon²

¹Department of Electrical and Computer Engineering

²Department of Computer Science

North Carolina A&T State University

Greensboro, NC 27411

Tel: (336) 334-7760 x 219 Fax: (336)334-7244

Abstract

This paper examines a new robust color scheme and an adaptive object tracking technique. There are several popular color schemes used in face tracking which include Normalized RGB, Hue, Saturation, and Hybrid type of colors. Hybrid color schemes provide improved results compared to any single color scheme technique. Extensive experiments show the new robust Hybrid color scheme produced superior results in various lighting conditions. In conjunction with the robust hybrid color scheme to track head movements a supporting algorithm was needed to approximate the random path of the head movement. Kalman filter is a famous estimation technique in many areas to predict the route of moving object. We tested and developed a random-walk Kalman filter to track unpredictable and fast moving objects. The random-walk Kalman filter tolerates for tracking of quick random movements made by a person, which was not accommodated by linear tracking techniques.

1. Introduction

For many computer vision applications, such as automatic speech recognition, 3D animation, and surveillance a robust and reliable automatic head tracking technique in various unmodified environments is vital. Recent research in this area shows great progress and promise. There are many approaches to track the head position on an image sequence. Some tracking modules are based on feature invariant, which is used to find out a structural feature, some are based on template matching, which is using a stored pattern to track head position (pattern can be 2D or 3D). Others include appearance-based method, which is using a trained model from a set of images to capture the representative variability of facial appearance. In this paper we explore a combination of a hybrid color scheme

module and a random-walk Kalman filter to track random head movement in a variety of environments.

Many researchers have exploited the relative uniqueness of skin color to track faces. Human skin color has been used and proven to be an effective feature in many applications. A weakness of these systems is their heavy reliance upon skin color that forbids skin-colored objects in the background and, more importantly, forbids the subject from turning around so that the back of his head, rather than this face, is visible [1].

Color image histogram is an effective method for the purpose of object recognition, segmentation or tracking. Color histograms are relatively invariant to many complicated, non-rigid motions like translation, rotation about the imaging axis, small off-axis rotations, scale changes and partial occlusion. The color histogram percentile features are useful to recognize the pattern of human face with relatively low complexity. Many methods have been proposed to build a skin color model. In this paper we proposed a new Hybrid color scheme with the support of additional Hue and Saturation analysis features that provide noticeable improvement in performance in various lighting conditions.

The Kalman filter is an optimal estimator to predict the next position of a moving object. It addresses the general problem of trying to estimate parameters of interest from indirect, inaccurate and uncertain measurements. However, general purpose of Kalman filter is only working well under slight movement and gradual speed on the image sequence. We need adaptive methods to overcome this problem.

Section 2 will cover the color performance analysis in head tracking to show the improved result of our new color scheme compared to

result of other systems. Section 3 covers random-walk Kalman filter to trace correct location of unpredicted and rapidly moving object. Finally, section 4 will provide conclusion of experiment result.

2. Analysis of Color Scheme for Head Tracking

In the RGB model, a color is expressed in terms that define the amounts of Red, Green and Blue light it contains. Normalized color space is a popular color representation to specify human skin color patterns. Since under normal lighting conditions the brightness of the face is not important for characterizing skin colors, we can represent skin-color in the chromatic color space. Chromatic colors, known as "pure" colors in the absence of brightness, are defined by a normalization process [2].

$$C_r = R / (R + G + B)$$

$$C_b = B / (R + G + B)$$

Even though the most common way of representing color is through the RGB color space. In this paper we can see this color model is quite sensitive to lighting conditions since the color attribute is combined with the brightness. Hue (color) component can be used for facial region localization because it is comparatively insensitive to illumination changes. Hue image is obtained by logarithmic color-space transform, RGB to HSV. However, simple Hue image can be easily affected by complex background texture. Additional Saturation component can compensate this lack of robustness to the intricate environment.

S. Birchfield [2] introduced his own color scheme; in our experiments we call it *the Stanford scheme*, which uses color space consisting of scaled versions of the three axes $B-G$, $G-R$, and $B+G+R$. The first two contain the chrominance information and are sampled into eight bins each, while the last one contains the luminance information and is sampled more coarsely into four bins. The big difference in his method is that he also considers luminance information. By using this scheme we could get fairly good tracking result. However, this scheme shows partial dependency on light condition and background texture.

We attempted to find a new color scheme that is robust enough for various light and background conditions. From our previous experiment, *Stanford scheme* showed a better result compared to other methods. But in addition to this scheme, the characteristic of insensitivity to illumination is required for a practical and dependable tracking module. A new Hybrid color scheme that utilizes additional Hue and Saturation features is the one we chose to achieve this goal.

The research was executed with various sequences of images under different light condition, background, and persons. For the objective comparison of result, all of four sequences were obtained from Vision lab website of Stanford University. Person in a sequence is always inside of frame by controlling the camera movement. These sequences include different races, light condition and background. Importantly, linear prediction technique was exploited to predict next head position for this test.

Table 1 and 2 shows head tracking result of various color schemes we chose for test. As it is shown below, Hybrid color histogram with (20(Stanford) + 4(Hue) + 4(Saturation)) bins gives the best results compared to Hue (16), Hue and Saturation (8 + 8), Normalized RGB, *Stanford scheme* (20) and Hue-hybrid (20 + 8(Hue)) color histogram.

We employed the average distance from the true center (Table 1) and the average success rate (Table 2) as performance measurements. True center of each frame was firstly obtained by manual operation through the whole sequence. Average distance was calculated based on this series of true center points. Each test was implemented both of X and Y directions to provide a better benchmark of tracking result evaluation.

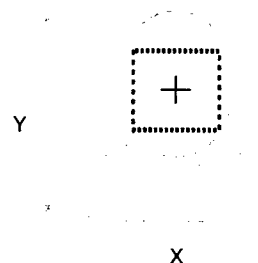


Figure 1: Manually grabbed facial region

Figure 1 shows the facial area and center point of that region. Hit number for each sequence of Table 2 is counted up when the destination point is located inside of this rectangular region. There is acceptable error range of five to ten pixels depends on the image.

From the result of Table 1 and Table 2, Hybrid color (20+4+4) gives 5.86 pixels distance to the X axis and 8.96 pixels to the Y axis. This is fairly good result compared to other two competent color schemes of Hybrid (20+8) and Stanford's (20). The result of Table 2 well supports this consequence.

We can expect better result only with additional Hue color (20+8). However, this color gave worse result for the sequence 3. Success ratio to the Y axis of sequence 3 is less than 50%. This means that Hue information is not stable enough to support Stanford color completely.

Stanford color scheme includes Normalized color and Regular RGB color. Even though their color system provides comparatively good results, it is still not robust enough under different conditions. Our test result shows that additional Hue and Saturation color features can attenuate the performance limitation of Stanford color.

3. Random-walk Kalman Filter

A robust head tracking requires a reliable prediction module for the estimation of the of the random moving objects. Our approach is on the base of Stan Birchfield's [2] method, which using intensity gradients, color histograms, and simple linear prediction. In gradient, an ellipse template is used to calculate the total gradient value around this ellipse within a suitable search window and then acquires a maximum value. In color, a face color histogram model will be created and used to match within the above search window. Birchfield also used a linear prediction to predict the search window on the oncoming frame according to the position of the previous 2 frames.

The main problem of the Birchfield method is the lack of accuracy if the moving speed of the head is too fast or the frame rate is too low. The result is a unreliable prediction window and the head position will be distracted. In this case, the way to improve the tracking performance is to increase the search range of search window,

however this will cause the processing speed down. So, there exists a limitation in using the linear prediction algorithms used by Birchfield.

In order to overcome this problem, we propose a random walk Kalman filter to predict the search window with a center of head position and a suitable range on the consecutive frames, and then update this prediction using the measurement value of the tracking head.

Kalman filter is an optimal estimator. It addresses the general problem of trying to estimate parameters of interest from indirect, inaccurate and uncertain measurements. Due to its recursion, new measurement data can be fed back to system as they arrive, so it can be used in real-time image processing system.

Kalman filter estimates a process by using a form of feedback control: the filter estimates the process state at some time and then obtains feedback in the form of (noisy) measurements. As such, the equations for the Kalman filter fall into two groups: time update equations and measurement update equations [4]. The time update equations are responsible for projecting forward (in time) the current state and error covariance estimates to obtain the *a priori* estimates for the next time step. The measurement update equations are responsible for the feedback—i.e. for incorporating a new measurement into the *a priori* estimate to obtain an improved *a posteriori* estimate. To adapt this prediction method to our random tracking needs we introduce new algorithms.

In our system, we construct the system model as random walk. Some related equations are as follows:

The state vector $\hat{x}_k = [x_k, y_k]$, where x_k, y_k indicate the center position of head on the k th frame image.

The measurement vector $z_k = [x_{z_k}, y_{z_k}]$, where x_{z_k}, y_{z_k} express the measurement value from our approach.

$$(1) \quad \hat{x}_k^- = u(t),$$

$u(t)$ = unity Gaussian white noise, that is random walk which means it has zero mean and unity variance [3].

$$(2) \quad z_k = Hx_k + v_k$$

From (1), (2), we can construct parameters of Kalman filter as follow:

Transimtion matrix

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, R = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix},$$

The initial *a priori* estimate error

$$P_0^- = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

It will show different performance by using different frame rate sequence of image. We captured some different image sequence with different frame rate, 10,24 frames per second. If we use 24 fps image sequence, there are no problems. Following sample results are from a 10 fps image sequence. In this sequence, the maximum head displacement between 2 consecutive frames is about 62 pixels. If using the linear prediction, the center of search window on the next frame would be out of tracking, particularly on turnover motion. That means it can't get the good result. However, we got good results in our approach using random walk Kalman filter. Figure 2 (a) and (b) show our experiment result of head tracking by using random walk Kalman filter.

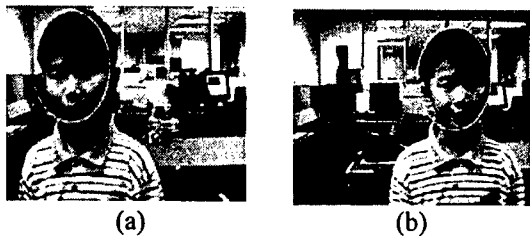


Figure 2 : Sample results from a 10 fps image sequence

Figure 3 shows the x-coordinate comparison of head position of Kalman filter, Birchfield's, and true center. The real head positions are recorded manually. There are several pixels calibration between Kalman filter and Birchfield's approach.

4. Conclusion

This paper presents a robust automatic visual tracking module that utilizes a new Hybrid color scheme with hue and saturation support and

random-walk Kalman filter for the prediction of the head. From our test result, we can conclude that proper mixture of all of RGB, chromatic color, Hue, and Saturation gives the best result compared with other currently available color schemes to track the human face. Moreover, if it can be combined with random-walk Kalman filter, the resulting module should provide a robust and reliable tracking method that overcomes many current problems in predicting the correct position of random and fast moving objects. The improvements in these two modules shows great promise for the development of a robust head tracking for ASR and other computer vision applications.

References

- [1] J. Yang, A. Waibel. A real-time face tracker. In *Proc. of WACV'96*, 1996, pp. 142-147.
- [2] Stan Birchfield. Elliptical Head Tracking Using Intensity Gradients and Color Histograms, In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, California, pages 232-237, June 1998.
- [3] Robert G. Brown and Patrick Y. C. Hwang, "Introduction to Random Signals and Applied Kalman Filtering", Second Edition, John Wiley & Sons, WC, p273-274, 1992.
- [4] "Open Source Computer Vision Library Reference Manual", Intel Corporation, Chapter 19, Copyright © 1999-2001

Table 1 : Average Distance from the True Center (unit : pixel)

	Seq. 1		Seq. 2		Seq. 3		Seq. 4		Avg. (pixel)	
Hybrid (20+4+4)	5.49	8.49	7.44	5.98	5.31	9.9	5.19	11.46	5.86	8.96
Hybrid (20+8)	4.49	8.49	8.5	7.03	16.09	17.45	3.38	8.17	8.12	10.29
Stanford	16.99	9.61	8.72	11.49	3.52	10.08	3.4	7.82	8.16	9.75
Hue+Saturation	23.86	21.51	15.05	14.28	3.15	9.56	6.44	9.13	12.13	13.62
Hue	33.56	20.29	13.7	12.31	9.3	10.88	7.65	10.18	16.05	13.42
Normalized	25.21	9.89	34.13	36.8	30.83	17.82	4.85	10.97	23.76	18.87
	X	Y	X	Y	X	Y	X	Y	X	Y

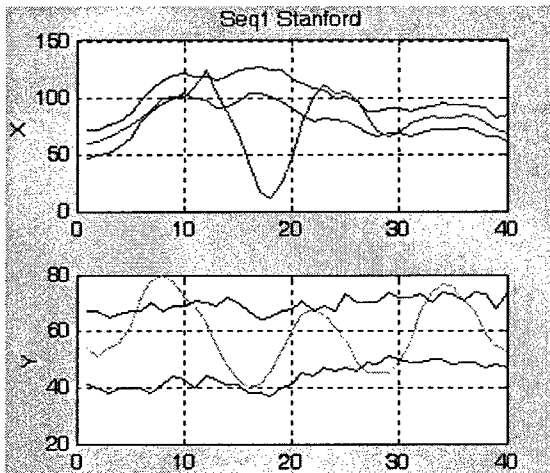
X : x direction tracking result

Y : y direction tracking result

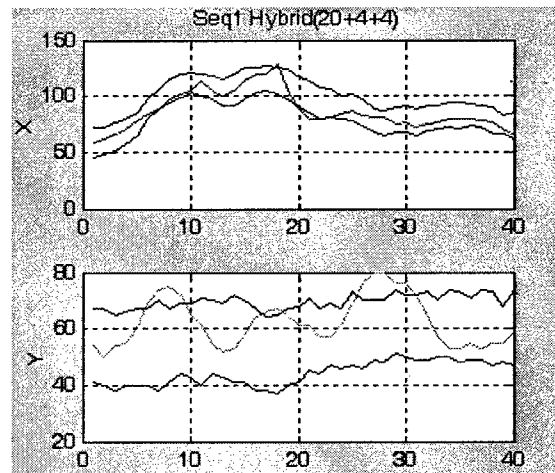
Table2 : Average Success Rate (Possibility to stay in the facial region through the whole sequence)

	Seq. 1 (40*)		Seq. 2 (65)		Seq. 3 (85)		Seq. 4 (101)		Avg. (%)	
Hybrid (20+4+4)	37	29	59	61	80	61	93	77	92.4	78.4
Hybrid (20+8)	39	33	51	59	59	40	101	97	85.9	78.7
Stanford	25	27	47	49	82	56	101	98	87.6	79.0
Hue+Saturation	14	18	36	35	81	62	91	81	76.3	67.4
Hue	14	21	33	39	62	49	84	80	66.3	64.9
Normalized	19	27	20	17	46	30	94	85	61.5	54.6
	X	Y	X	Y	X	Y	X	Y	X	Y

* : # of frames in a video sequence



(a)



(b)

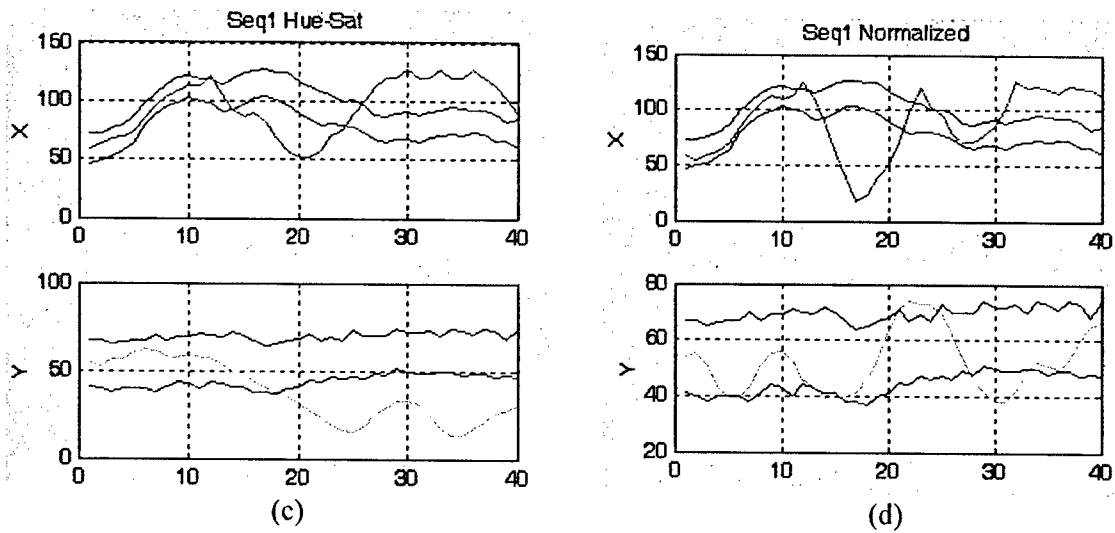


Figure 1 : (a) Stanford $(B-G)+(G-R)+(R+G+B/3)$ (b) Stanford + Hue(4) + Saturation(4)
(c) Hue + Saturation Color Scheme (d) Normalized Color

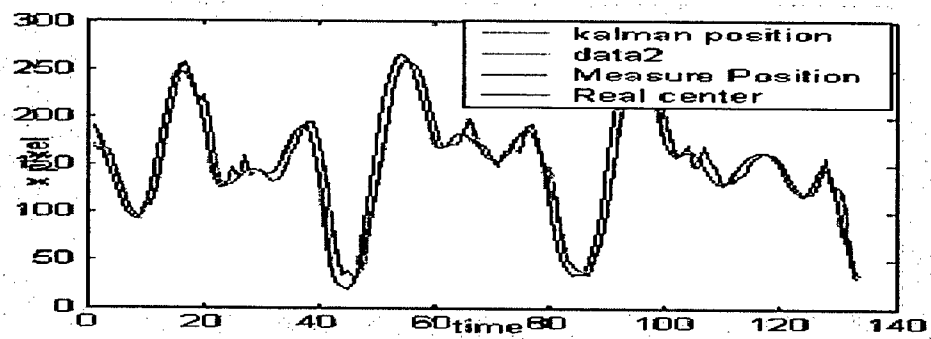


Figure 3 : Comparison x-coordinates of head position with Kalman filter, Birchfield, and real center position (manually recorded).

The Validation of Military Callsign Intelligibility

Celestine A. Ntuen & Misty Blue

The Institute for Human-Machine Studies

Department of Industrial & Systems Engineering

North Carolina A&T State University

Abstract

This study was conducted to evaluate the performance of human perception of speech generated by computers under normal and stressful military environments. Performance intensity (PI) functions for speech intelligibility were developed. Results are used to determine human speech awareness thresholds (SAT) for quiet and noise environments.

1. INTRODUCTION

Our ability to perform tasks effectively in environments such as the battlefield, airspace management (pilots and air traffic controllers), hospitals, and manufacturing systems, depend in part on our ability to process speech signals. Effective speech communication requires clear speaking by the talker, nonrestrictive transmission channel (medium), and good hearing and speech comprehension by the listener. These capabilities have been tested using various speech material and trained takers (speech understanding tests) or listeners (speech intelligibility tests).

One of the several methods to measure our ability to process information generated by sound or speech signals is known as speech intelligibility (Logan, Greene, & Pisoni, 1989).

Speech Intelligibility (SI) is an index for measuring the minimum absolute threshold of perceiving sound in a given environment. SI is quantitatively defined as the percentage of speech units that can be correctly identified by a listener over a given communication system in a given acoustic environment or the degree to which speech can be understood during given conditions (Letowski, Karsh, Vause, Shilling, Ballas, Brungart & McKinley, 2001). Intelligibility tests evaluate the number of words or other speech units that can be correctly identified within a controlled situation. Some examples of speech intelligibility tests are documented in ISO (1986). The relevant ones to this study are:

Diagnostic Rhyme Test (DRT): The DRT uses a set of isolated words to test for consonant intelligibility in initial position (Goldstein, 1995; Logan, Greene & Pisoni, 1989). The tests consist of 96 word pairs that differ by a single acoustic feature in the initial consonant. Word pairs are chosen to evaluate the phonetic characteristics.

Modified Rhyme Test (MRT): The MRT is an extension of DRT, tests for both initial and final consonant apprehension (Logan, Greene & Pisoni, 1989¹). The test consists of 50 sets of 6 one-syllable words that make a total set of 300 words. The set of 6 words is played one at the time and the listener marks which word he thinks he hears on a multiple choice answer sheet.

Diagnostic Medial Consonant Test (DMCT): The DMCT is the same type of test as the rhyme tests described before. The material consists of 96 bi-syllable word pairs like "stopper-stocker" which were selected to differ only with their intervocalic consonant.

2. MILITARY CALLSIGN TEST (CAT)

The Auditory Research Team at the United States Army Research Laboratory developed the CAT test (Letowski, Karsh, Vause, Shilling, Ballas, Brungart, & McKinley, 2001). The CAT test utilizes military callsigns for calling phrase. A single callsign for CAT consists of a word and a number. The word is a two-syllable military alphabet code and a one-syllable number, for example, alpha 1 or bravo 2. due to their familiarity with test material and task environments. To maintain its ecological

Multi-modal Speech Recognition Workshop, June 10-12, 2002/NCAT--
Greensboro

validity, it is important to test the CAT in quiet conditions so as to establish a standard and a reference SI metric for comparison with other standard SI metrics (ISO 1986). The test material seems to be a good compromise between (1) simplicity and poor predictive value of monosyllabic signals and (2) complexity and memory load of nonsense sentences and long number sequences (Letowski, 2001).

The CAT test has been informally used by the ARL-ART in several studies but is still lacking proper validation and standardization. Such a process requires several steps that need to be completed before the final version of the test may be released. One of these steps is the standardization of SI and evaluation of the related performance intensity (PI) curve for CAT both in quiet and with background noise

3. PROCEDURE & METHODOLOGY

Participants

A group of 24 listeners between the ages of 18 and 45 participated. All listeners

The listeners repeated the test with signal level increasing in 5dB steps until they achieve 95% or better on both tests (RMS and PEAK recordings). All the listeners' responses were stored in a file and subsequently imported into an Excel™

4. SAMPLE RESULTS

had pure-tone hearing thresholds better than or equal to 20dBHL at audiometric frequencies from 250Hz through 8000Hz (ANSI S3.6-1996) and no history of otologic pathology. An audiometric screening test was performed prior to participation in the study.

Each listener was seated at the listener station in a sound treated test booth using an IBM PC/586 computer and wearing TDH-39 testing earphones. All the instructions were displayed on the computer screen and the participant was able to use either the computer mouse or the computer keyboard for data input. The listener was asked to listen to the series of the CAT (military alphabet callsigns and one syllable numbers 1-8) items and identify them by pressing appropriate keys on the computer keyboard. Also, the main screen showed the display CAT test (Peak or RMS) and the signal-to-noise ratio (SNR) given by -18 dB, -12dB, -8dB, 0dB, 6dB, 12dB.

spreadsheet for analysis. Each listener participated in a single listening session. The session lasted about four hours and included audiometric screening, instructions, testing and several 10-15 minute long breaks.

The PI function showed some characteristics of logistics distributions (See example in Figure 2).

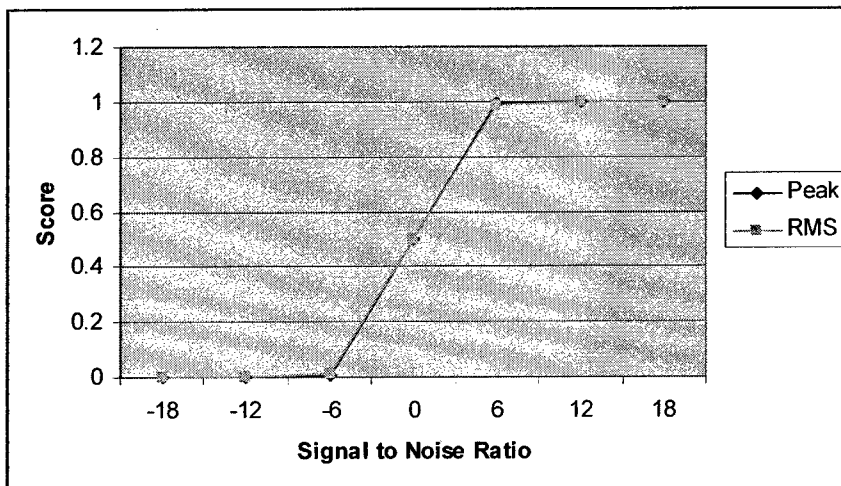


Figure 2: Sample logistics PI function for CAT intelligibility

$$\text{Score} = \frac{1}{1 + e^{-0.78235 \cdot \text{SNR}}} ; R^2 = 90\%$$

(Peak) (1)
 $0 < \text{SNR} \leq 11.77$

$$\text{Score} = \frac{1}{1 + e^{-0.745 \cdot \text{SNR}}} ; R^2 = 88.24\%$$

(RMS) (2)
 $0 < \text{SNR} \leq 12.36$

Figure 2: Sample logistics PI function for CAT intelligibility

5. CONCLUSION

The logistics PI models show that speech awareness threshold (SAT) occurs at signal-to-noise-ratio (SNR) > 0, with the average listener achieving an SI value of 95% at SNR values of 11.64 for Peak and 12.22 for RMS. By using simple one parameter linear model, speech awareness threshold occurs at SNR values of approximately 2 for both Peak and RMS tests, with the average listener achieving an SI value of 95% at SNR values between 7.7 and 7.9.

References

Goldstein, M. (1995). Classification of methods used for assessment of text-to-speech systems

according to the demands placed on the listener. *Speech Communication*, 16, 225-244.

Jekosch U. (1993). Speech quality assessment and evaluation. *Proceedings of Eurospeech 93* (2): 1387-1394.

Letowski, T., Karsh R., Vause, N., Shilling, R., Ballas, J., Brungart, D., McKinley, R. (2001).

Human Factors Military Lexicon: Auditory Displays. Aberdeen Proving Grounds, MD.

Letowski, T. (2001). Performance Intensity function for the Callsign Acquisition Test (CAT)

Research Protocol. Aberdeen Proving Grounds, MD.

Logan J., Greene B., Pisoni D. (1989). Segmental intelligibility of synthetic speech produced by

rule. *Journal of the Acoustical Society of America*, JASA. 86 (2): 566-581.

Large Vocabulary Audio-Visual Speech Recognition

Chalapathy Neri & Gerasimos Potamianos
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

- Motivation
- A/V speech recognition architecture
- Visual feature extraction
- Audio-visual fusion
- Results
- Challenges and conclusions

Pervasive Speech recognition

➤ Pervasive deployment of speech will require better recognition in degraded acoustic conditions:

- High noise ("speech babble") e.g.

- ✓ Military applications
- ✓ Automobiles
- ✓ Video Games & Interactive television



- Whispered Speech

- Privacy

- Lombard speech

- High-noise conditions



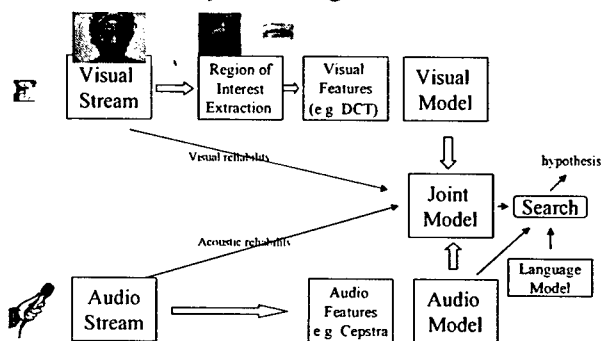
- Speech pathology

Audio-Visual speech recognition is a key enabler

IBM's A/V speech effort

- History (www.research.ibm.com/AVSTG)
 - About a 3 year old effort
 - Led the JHU Workshop team on A/V speech recognition, 2000
 - AVSTG department formed in 2001
 - Taught an invited ELSNET tutorial on A/V speech recognition (Prague, 2001)
- Highlights/differentiators of our work
 - One of a kind database for AV LVCSR
 - State-of-the-art audio ASR subsystem (LVCSR)
 - Fully automated visual front end
 - Multiresolution face detection
 - Augmented visual speech ROI (jaw region instead of mouth)
 - Multistage (linear transform based) visual feature extraction
 - Sub-phonetic visual speech models
 - Scales to large-vocabulary recognition
 - Phone-level asynchronous A/V fusion
 - Joint a/v model training
 - Maximum entropy based stream weight estimation (global and local)
 - Multiple domain exploration
 - Read speech (digit/C&V/LVCSR), Impaired speech, Automobile, Broadcast News
 - Visual adaptation to new domains

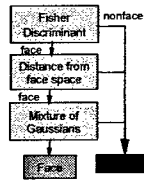
Audio-visual speech recognition: architecture



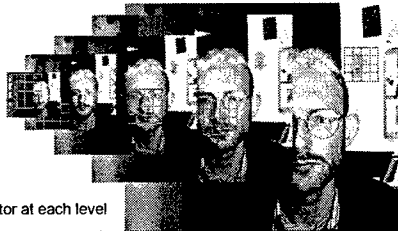
Visual Front-End

Multiresolution face detection:

- Search for skin-tone pixels
- Search image pyramid across scales & locations.
- Each square $m \times m$ region is considered as a face.
- Hierarchical, pixel based approach, using LDA and PCA.



5-level pyramid with face detector at each level



ROI EXTRACTION

- Steps (after face/mouth tracking):
 - Smooth mouth center and size estimates by median filtering.
 - Extract a 64×64 pixel, size-normalized mouth ROI.
 - ROI includes jaw and cheeks.



VISUAL FEATURES

- Three stages:

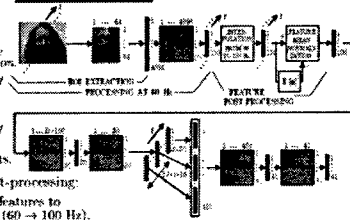
I: Discrete cosine transform (DCT) for data compression.

II: Intra-frame LDA / MLT: $100 \rightarrow 30$ static features.

III: Inter-frame LDA / MLT: $15 \times 30 \rightarrow 41$ dynamic feats.

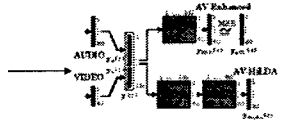
- DCT feature post-processing:

- Interpolate visual features to audio feature rate ($60 \rightarrow 100$ Hz).
- Apply feature mean normalization for lighting compensation.

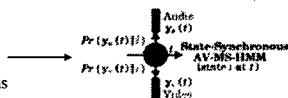


Audio-visual Fusion Techniques

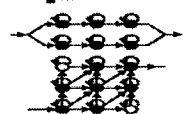
- Feature fusion
 - Enhancement approach
 - Discriminant fusion (HiLDA)



- State Synchronous Multistream HMM
 - Allows weighting decisions



- Phone synchronous Multistream HMM
 - Allows asynchrony within a phone



THE MULTI-STREAM HMM FOR AV-ASR

- The multi-stream (MS) HMM:
 - Observation conditional "score" of audio-visual state $i \rightarrow (i_a, i_v)$:

$$\mathcal{L}(y(t)|i) = [Pr(y_a(t)|i_a)]^{\lambda_a} \times [Pr(y_v(t)|i_v)]^{\lambda_v}$$
 - Exponents model stream "reliability". Typically:

$$0 \leq \lambda_a, \lambda_v \leq 1, \lambda_a + \lambda_v = 1$$
- State vs. phone level synchronous (product) MS-HMM:
 - State **synchrony**: $\{i_a\} = \{i_v\}$, $i = i_a = i_v$.
 - State **asynchrony**: $i \in \{i_a\} \times \{i_v\}$.
- MS-HMM parameter training:
 - Model parameters: $\theta = \{\theta_a, \theta_v, \lambda_a, \lambda_v\}$, where θ_a, θ_v are audio- or visual-only HMM stream params.
 - Maximum likelihood estimation (MLE) of θ_a, θ_v via EM:
 - Independent E- and M-steps for MLEs of θ_a, θ_v .
 - Joint audio-visual MS-HMM E-step, M-step, as above.
 - MLE of λ_a, λ_v is impossible. Instead, we have considered:
 - Parameter grid search. Minimizes held-out data WER.
 - Minimum classification error (MCE) training by GPD.
 - Maximum entropy. Maximizes data posterior log-likelihood.

IBM VVAV databases

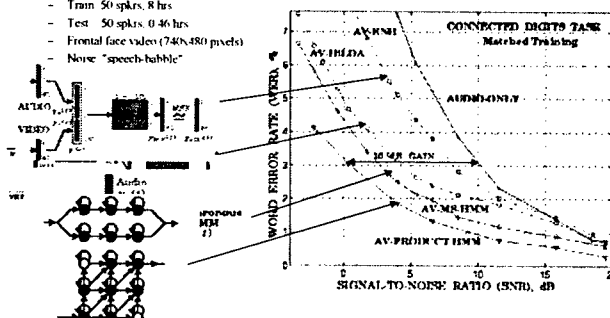
- LVCSR
 - First-of-a-kind audio-visual database for large-vocabulary continuous SI speech recognition (LVCSR)
 - 290 subjects
 - 70 hrs. continuous speech, 10,400 word vocabulary
- Digits
 - 50 subjects
 - 8 46 hrs. continuous speech, 11 word vocabulary
- Database Format
 - Frontal face color video, 704x480, 30 Hz, MPEG2
 - 16 kHz/16bit pcm



Experiments on Digits

Fusion Techniques

- IBM VVAV database digits
 - Train: 50 spkrs, 8 hrs
 - Test: 50 spkrs, 0.46 hrs
 - Frontal face videos (740x480 pixels)
 - Noise: "speech-bubble"



Results - Summary

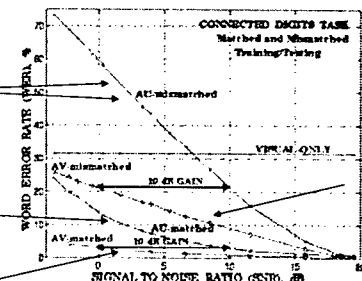
- Effective gain of 10 dB @ 10 dB SNR (relative to mismatched audio)
- Effective gain of 10 dB @ 10 dB SNR (relative to matched audio)

Digits Task

Train in clean
Test in noise

Train in noise
Test in noise

Matched
Audio + visual

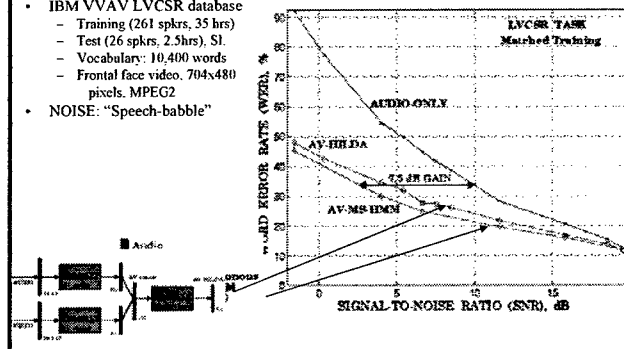


Mismatched
Audio + visual

Experiments on LVCSR

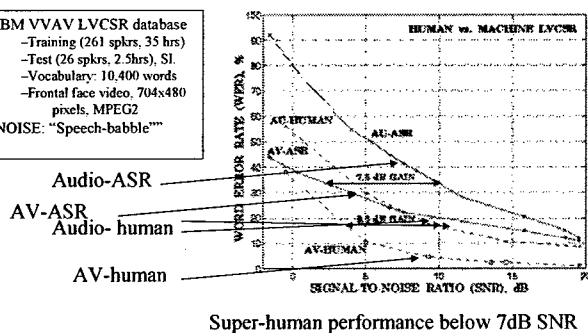
Results: LVCSR

- IBM VVAV LVCSR database
 - Training (261 spkrs, 35 hrs)
 - Test (26 spkrs, 2.5hrs), SI
 - Vocabulary: 10,400 words
 - Frontal face video, 704x480 pixels, MPEG2
- NOISE: "Speech-babble"



Results: Human vs. Machine

- IBM VVAV LVCSR database
 - Training (261 spkrs, 35 hrs)
 - Test (26 spkrs, 2.5hrs), SI
 - Vocabulary: 10,400 words
 - Frontal face video, 704x480 pixels, MPEG2
- NOISE: "Speech-babble"



Challenges

- IBM VVAV data
 - Audio
 - Read Speech, single microphone
 - Additive "speech babble" noise
 - Video
 - Frontal Face, uniform background
 - Uniform lighting
- Broadcast Video data
 - Audio
 - Spontaneous speech
 - Additive music, varying channel, etc.
 - Video
 - Limited pose variation, background clutter
 - Uniform lighting
- Automobile data
 - Audio
 - Spontaneous speech
 - Automobile noise (speed variation, radio, seat belt, etc.)
 - Video
 - Pose variation, varying background
 - Non-uniform lighting Conditions

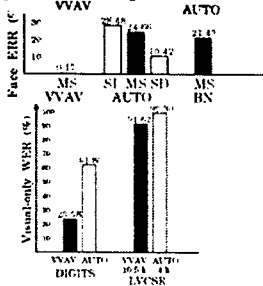


VISUALLY CHALLENGING DOMAINS: PRELIMINARY RESULTS

- Challenge: Video data variability in *head pose*, *background*, and *lighting* affects *face detection*, *ROI extraction/normalization*, thus visual- and AV-ASR.

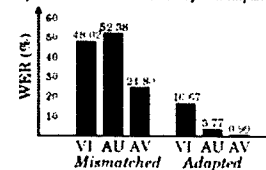


- Face detection error for VVAV, AUTO (multi-speaker vs. speaker-independent) and RN data
- Visual-only WER for VVAV vs. AUTO domains (DIGITS and LVCSR tasks)



AUDIO-VISUAL SPEAKER ADAPTATION

- Important for speaker enrollment and limited data domains, but hardly ever considered in the AV-ASR literature
- Main techniques:
 - MLLR: Rapid adaptation of HMM stream component means
 - MAD: Bayesian approach, adapts all HMM parameters
 - FE: Front end adaptation of LDA/MLT matrices
- The domains/tasks considered:
 - Domains: Noisy trading floor; hearing impaired speech
 - Tasks: LVCSR, DIGITS
- MLLR adaptation results on DIGITS speech impaired data:



Conclusions

- Consistent and significant gains for all audio conditions
- Significant performance gains in "speech-babble" noise
 - Effective gain of 10 dB @ 10 dB SNR for digits
 - Effective gains of 7.5 dB @ 10 dB for LVCSR
- Significant gains in relatively clean environments
 - 62% relative gain for digits (0.75 -> 0.28)
 - 8% for LVCSR
- Super-human performance at high-noise levels
- Asynchrony modeling helps for digits
- Further research required in visually challenging domains
- Visual adaptation is a promising approach
 - Upto 67% relative improvement in visual speech recognition

Who ? What ? Where ? How ?

Perceptually Aware User Interfaces

Alex Waibel

June 11, 2002
Interactive Systems Laboratories
Carnegie Mellon University
University of Karlsruhe

<http://www.is.cs.cmu.edu>
Email: waibel@cs.cmu.edu

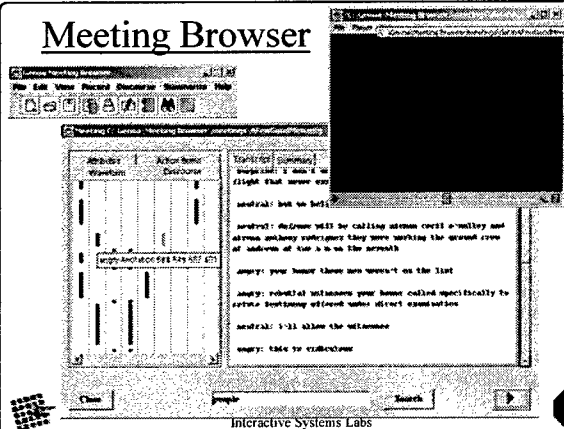
Interactive Systems Labs

Meetings



Interactive Systems Labs

Meeting Browser



Interactive Systems Labs

Interpreting Human Communication

"Why did Joe get angry at Bob about the budget ?"

Need Recognition and Understanding
of Multimodal Cues



- Verbal:
 - Speech
 - Words
 - Speakers
 - Emotion
 - Genre
 - Language
 - Summaries
 - Topic
 - Handwriting
- Visual
 - Identity
 - Gestures
 - Body-language
 - Track Face, Gaze, Pose
 - Facial Expressions
 - Focus of Attention

Interactive Systems Labs

Human Interaction

- People ID – Who?
 - Speaker ID, Face ID
 - Type: Dominant, Submissive, etc.
 - Relationship: Family, Friends, Colleague
- Speech and Discourse – What ?
 - Speech: Transcript
 - Discourse States (Speech Acts, Topics), Games, Turn Taking
 - Discourse Types and Genres (Negotiation, Chatting, Lecturing)
- Emotional State, Affect – How ?
 - Angry, Happy, Sad, Afraid:... Busy, Nervous, Relaxed
 - Discourse Style: Sloppy, Formal, Colloquial
- Localization and Speaker and Focus of Attention – Where ?
 - Speaker Localization
 - Focus of Attention Tracking



Interactive Systems Labs



Main Challenge and Goal

Robustness in Real-Life Situations



Interactive Systems Labs



Visual Challenges

Low quality



Illumination



Head pose



Occlusion



Interactive Systems Labs



Acoustic Challenges

- Sloppy Speech
- Noise
- Reverberation
- Acoustic Scene Analysis
- Cross Talk
- Distant Mic



Interactive Systems Labs



Where ?

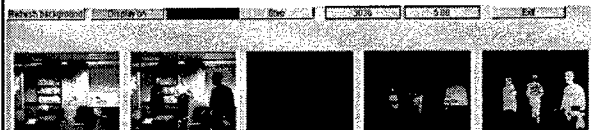
- Face Tracking (Visual)
- Sound Source Localization (Acoustic)
- People Tracking (Visual)
- Behavior and Movement Models



Interactive Systems Labs



Tracking People



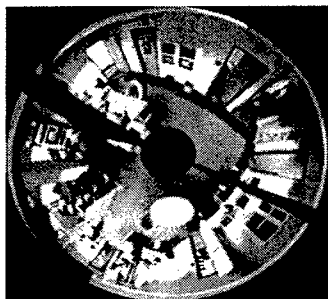
Tracking People



Interactive Systems Labs



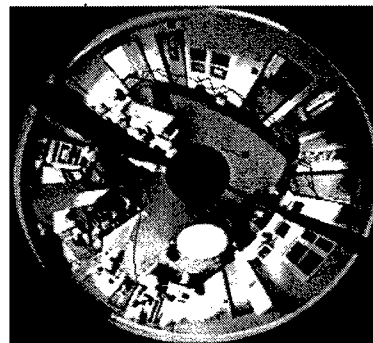
Tracking Multiple People



Interactive Systems Labs



From Tracking to Modeling Activity



Interactive Systems Labs



Real-Time Face Tracker

Three Types of Models have been employed

- skin-color model to register the face
- motion model to estimate image motion
- camera model to predict and compensate for camera motion (pan, tilt, zoom)

The Face Tracker

- tracks a persons face while person is freely moving (e.g. walks, jumps, sits down and stand up)
- Framerate : 30+ frames per second using a low end workstation (HP9000) or Pentium II 266 PC.



Interactive Systems Labs



Real-Time Face Tracker



Interactive Systems Labs



Using a Panoramic Camera



Cyclovision's ParaCam



Camera View

Panoramic View



Interactive Systems Labs



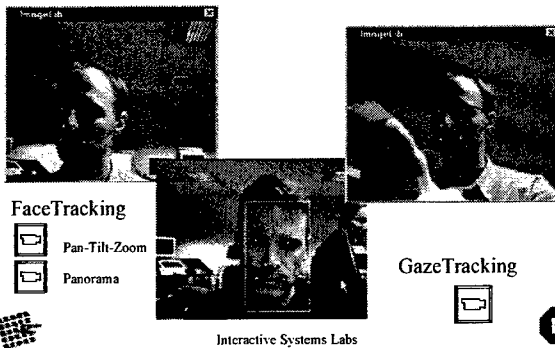
Pose Tracking by Modeling Shape



Interactive Systems Labs



Face and Pose Tracking



What ?

- Large Vocabulary Speech Recognition
 - Issues:
 - Sloppy Speech
 - Distant Microphones
 - Mismatch in Vocabulary
 - Other Languages
 - Many Other Aspects: Topic Detection, Named Entity, Translation, Discourse,
- Multimodal Dialog
 - Fuse Speech, Pointing, Gesture, Handwriting
 - Fusion Usually at Semantic Level
- Audio-Visual Speech
 - Combine Speech and Visual Info

Interactive Systems Labs

From Read Speech to Conversational Speech

- Wall Street Journal Dictation
- Broadcast News Database
 - Transcription and Information Retrieval on News Casts
 - Multilingual Speech Recognition
- Switchboard & Callhome
 - Human to Human Telephone Speech
- Meetings and Discussion Database
 - Newshour (18h)
 - Crossfire (9h)
 - Group Meetings

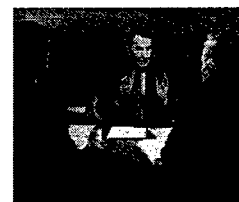
Conversational Speech

Interactive Systems Labs

Transcribing Speech in Meetings

Run-On Transcription of Meetings

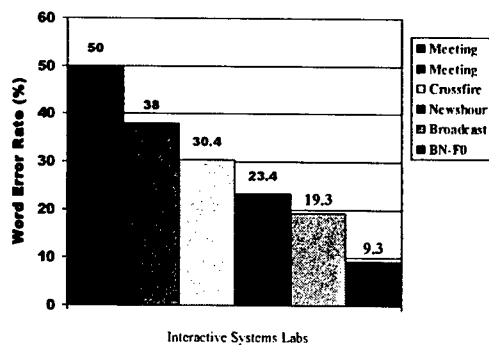
- *Mismatched* Recording Conditions
 - Remote Microphones
 - Cross-Talk
 - Recording Always on !
 - Noise
 - Multiple Speakers
- *Mismatched* Speaking Style:
 - Spontaneous and Conversational Human to Human Speech
 - Emotional Speech
- *Mismatched* Language and Vocabulary
 - Special Ideosyncratic Topic



- Three Tasks:
 - Newshour
 - Crossfire
 - Group Meetings

Interactive Systems Labs

Recognition of Conversational Speech

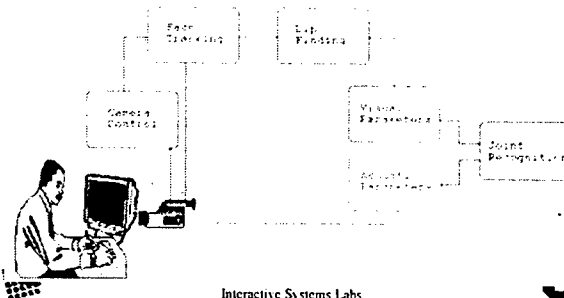


Interactive Systems Labs

Audio-Visual Speech:

(When Acoustic Processing is not Good Enough)

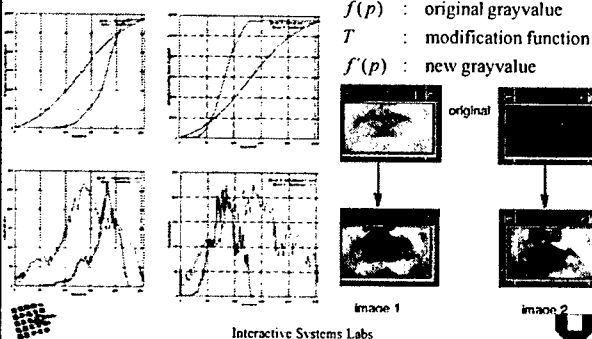
- Duchnowski, Manke, Bregler, Meier, Waibel
- ICASSP'93, ICSLP'94, ICASSP'95, ...



Interactive Systems Labs

Visual Preprocessing

grayvalue modification - example histogram : $f'(p) = T(f(p))$

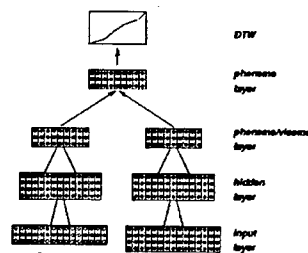


Interactive Systems Labs

Audio-Visual Recognizer

$$hyp_c = \lambda_a hyp_a + \lambda_v hyp_v$$

$$1 = \lambda_a + \lambda_v$$



Features

- What Features to Use?

Fusion Level

- Feature Vector
- Phone Streams
- Word Level

Fusion Methods

- Trained Weights
- Entropy Weights
- SNR Weights

Interactive Systems Labs

Experiments

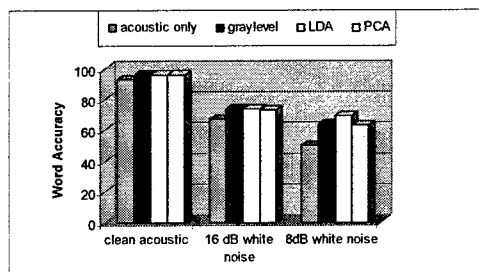
- Task: Continuous Letter Spelling
 - Difficult, but smaller Vocabulary
- Speaker dependent
 - audio-visual results
 - Fusion by Entropy Weights
 - LDA Front End
 - Phone Level Fusion



Interactive Systems Labs



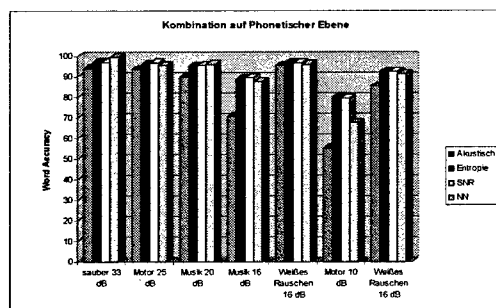
Audio-Visual Fusion



Interactive Systems Labs



Fusion Weights and Noise



Interactive Systems Labs



Who ?

- Once we have found the Face
 - Problems: Occlusion
- Face ID
 - Problems: Occlusion
- Speaker ID
 - Problems: Distant Mics, Noise, Jamming Noise
 - Phonetic Speaker ID, Qin Jing



Interactive Systems Labs



People Identification: Challenges

Low quality



Illumination



Head pose



Occlusion



Interactive Systems Labs



How ?

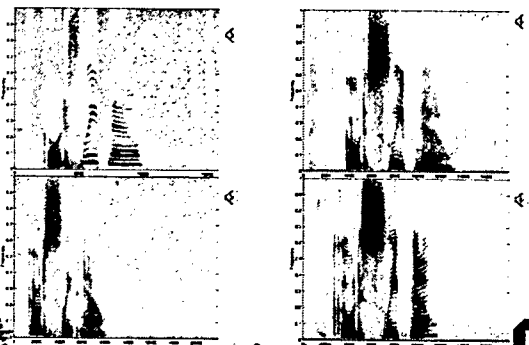
- Detect Emotional State
 - Happy, Angry, Sad, Afraid
 - Distress, Busy, Relaxed...
- Techniques:
 - Acoustic: (Polzin, 1999)
 - Prosody: Intensity, Pitch, Rhythm,
 - Language: Words and Expressions Used
 - Visual: (Cohen)
 - Facial Expressions



Interactive Systems Labs



Emotion: Acoustic Information



Interactive Systems Labs



Emotion: Language Information

- Lexical metaphors
Son of a bitch!
(As Good as it Gets)
- Connotation-loaded lexems
You're a spoiled rotten little brat!
(Kramer versus Kramer)
- Intensification
We ain't got the slightest f... idea what happened ...
(Reservoir Dogs)
That makes me very very mad ...
(The Sweet Hereafter)

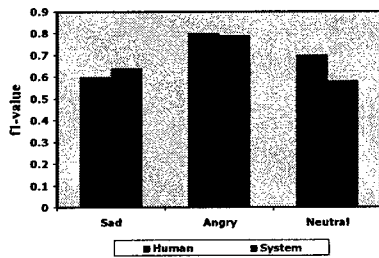


Interactive Systems Labs



Performance Comparison (Movies)

Verbal and Non-Verbal Information



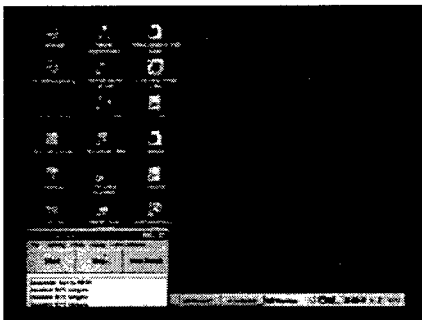
Interactive Systems Labs

To Whom ?

- Focus of Attention Tracking
 - R. Stiefelhagen, PUI'98, Humanoids'01, PhD Thesis'02
 - Who is addressee of an utterance ?
 - Who is someone making talking to ?
 - What is a human user attending to ?
- Observation:
 - FoA is a Psychological State, can only be inferred or 'guessed' from correlates
 - Both Observed User and Target are important:
 - Pose, Eye-Gaze
 - Possible Targets: Noise, Movement, Faces, Speech

Interactive Systems Labs

Focus of Attention Tracking



Interactive Systems Labs

Conclusion

- Complete Model of Human Communication is Needed
 - Include all modalities
 - Include different not only *what* was said, but also: *who, where, to whom, how...*
- Challenges:
 - Robust Processing of Component
 - Proper Level and Method of Fusion
 - Robust and *Dynamic* Fusion of Useful Clues

Interactive Systems Labs

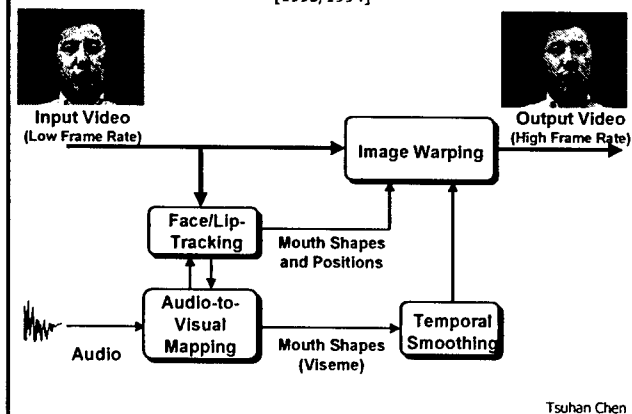
Joint Audio-Visual Speech Recognition and CMU Audio-Visual Speech Data Set

Prof. Tsuhan Chen
Carnegie Mellon University

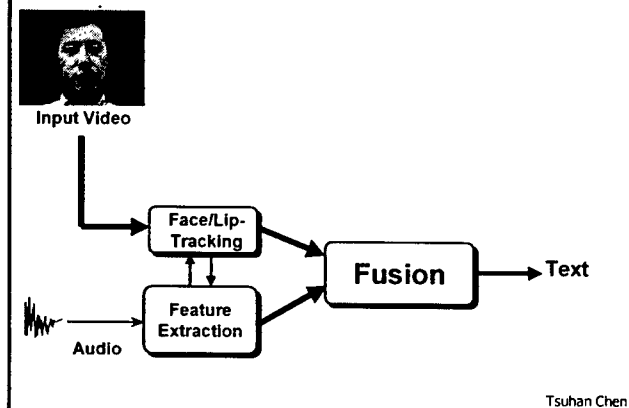
Thanks to Dr. Simon Lucey and Jie Huang

Where We Started...

[1993/1994]



Lip-Reading



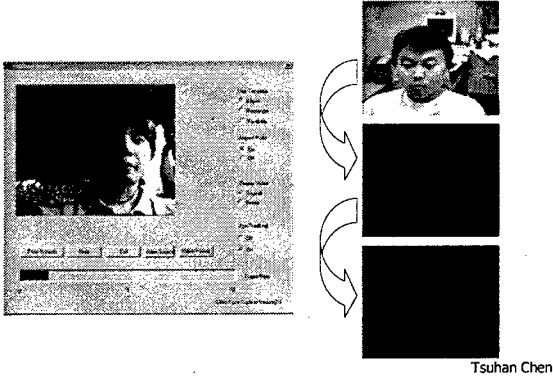
Audio-Visual Speech Data Set

- Thanks to Intel
- 78 isolated words 10 times
 - Date/time/month/day/etc.
 - Audio: 44.8 kHz, 16 bits
 - Video: 30/60Hz, 720x640
- Lip parameters extracted
- Noises
 - Gaussian white/pink noise, car, factory (Noise-X 92)
 - Babble/crosstalk
 - Lombard Effect



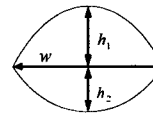
Tsuhan Chen

Face Tracking



Lip Tracking

- Modeling color distribution of mouth pixels
 - Gaussian mixture
- Deformable template



Tsuhun Chen

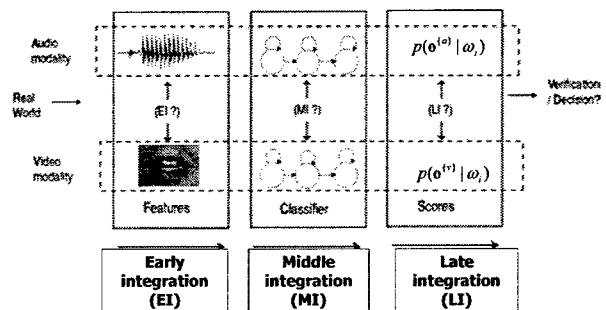
Customers...

- "Signal Processing for Media Integration," ICASSP 2002
 - Coupled HMM for Audio-Visual Speech Recognition, Nefian et al., Intel
 - Visual Speech Feature Extraction for Improved Speech Recognition, Zhang, Mersereau, Clements, Georgia Tech
 - Audio-Visual Speech Modeling Using Coupled HMM, Chu, Huang, UIUC
- Others

California State University	Queensland University of Technology
Chungghwa Telecom Lab, Taiwan	National Tsinghua University
DongYang University, Korea	National University of Singapore
Fabbrica Servizi Telematici, Italy	Norwegian Computing Center, Norway
IIT Bombay, India	Shanghai JiaoTong University, China
Instituto Tecnológico de Buenos Aires	Washington University
On2.com	

Tsuhun Chen

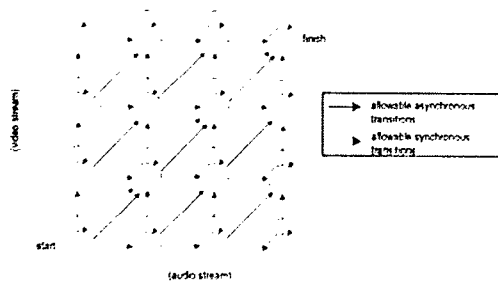
Fusion Techniques



Tsuhun Chen

Middle Integration (MI)

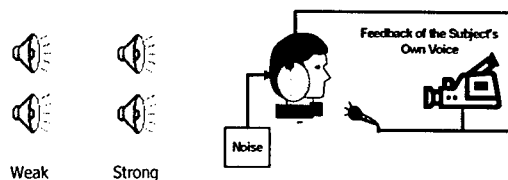
- Multistream HMM



Tsuhun Chen

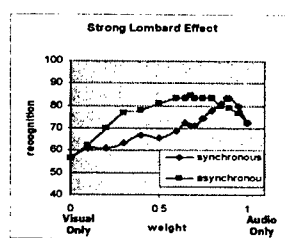
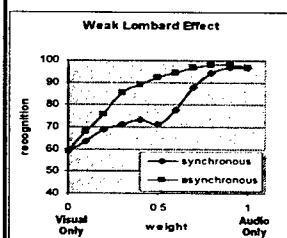
Lombard Effect

- Feedback
 - Voice changes with background noise
 - Lip movement changes too
- Data set



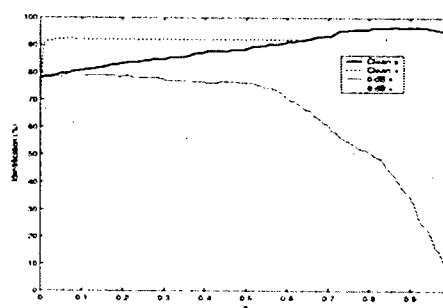
Tsuhun Chen

Result



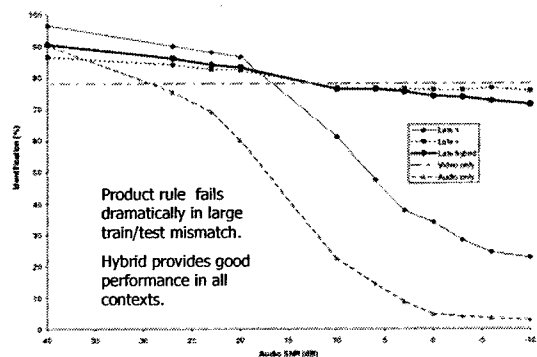
Tsuhun Chen

Product Rule vs. Sum Rule (For Speaker Identification)



Tsuhun Chen

Product Rule vs. Sum Rule



Tsuhun Chen

Quick Recap

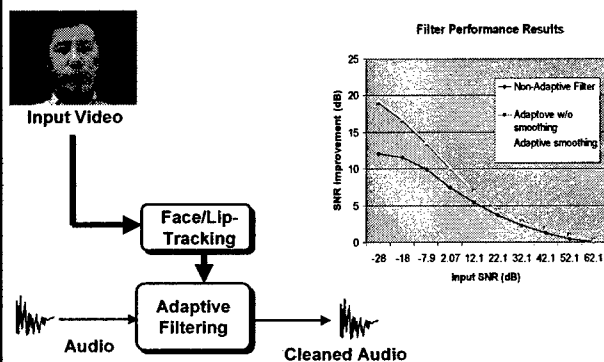
- Asynchronous MI (LI) has more freedom than synchronous MI (EI) → Better performance
- Product rule is better in Bayesian sense, but sum rule is more robust to mismatch
- Robustness to weighting
- Need to be careful about Lombard Effect
- Key to multimodal fusion
 - To model dependency between audio and visual signals
 - To dampen independent audio and visual noises

Tsuhun Chen

Beyond Multimodal ASR...

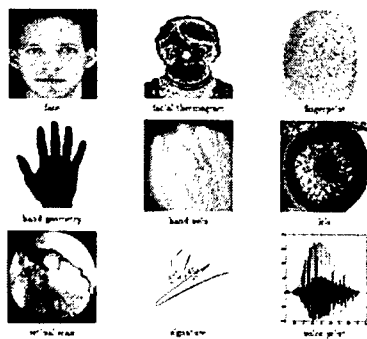
Tsuhun Chen

Visual-Assisted Speech Enhancement



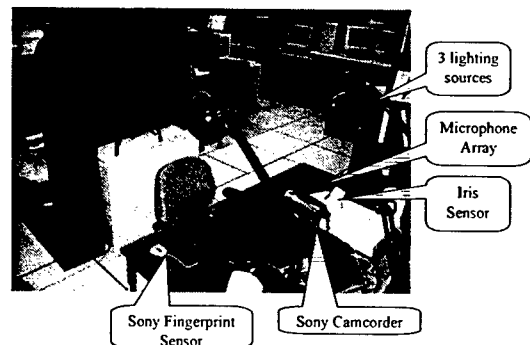
Tsuhun Chen

Multimodal Biometrics



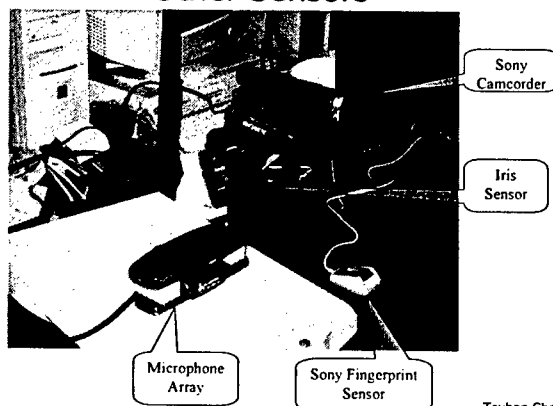
Tsuhan Chen

Data Collection



Tsuhan Chen

Other Sensors



Tsuhan Chen

CMU Multimodal Biometrics Database

Face:

- 30 subjects with 300 images each
- Image size: 720*480
- Different lighting conditions, with/without glasses and ambient lighting



Fingerprint:

- Image size: 192*128
- 50 images each finger



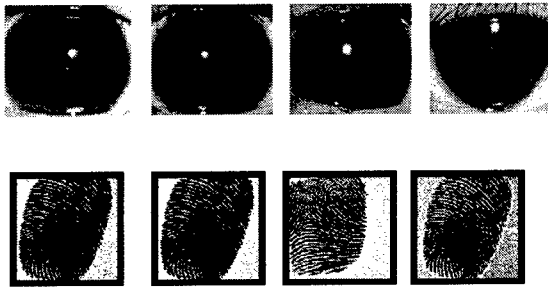
Iris

- Iris size: about 400*400
- 10 images each eye



Tsuhan Chen

Fingerprint and Iris Images



Tsuhun Chen

Multimodal User Interfaces

[CMU-GM Lab]

Face/Eye/Hand Tracking:
 . Driver-Vehicle Interfaces
 . Cognitive Overflow Study



Interview Video

Airbag Deployment Control
 Mirror/wheel/panel/seat adjustment



Driver ID and Encryption:
 Security, Safety, User Preference



Demo Vehicle



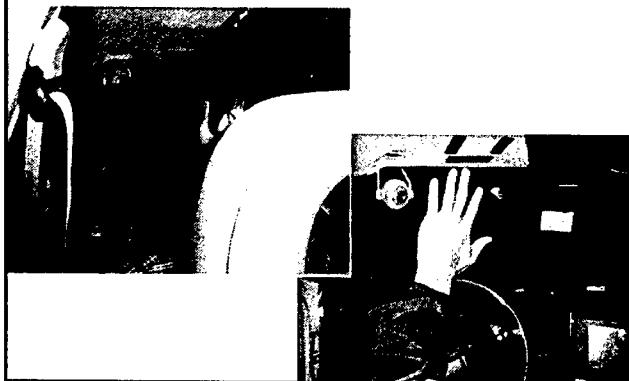
Tsuhun Chen

Demo Vehicle



Tsuhun Chen

FaceCam/GestureCam



"Visual is not noise-free"

Tsuhun Chen

Challenges...

Pose/Registration



Illumination

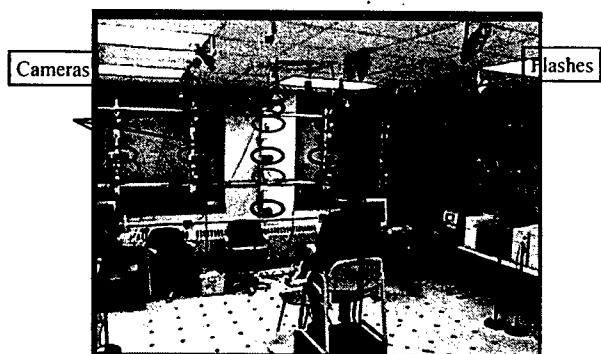


Expression



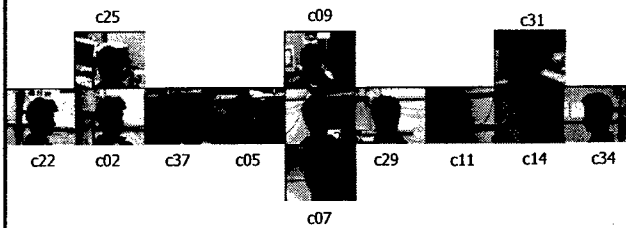
Tsuhun Chen

CMU PIE Database



Tsuhun Chen

Pose Variation



Tsuhao Chen

Illumination Variation

- 22 illumination conditions with background light



- 21 illumination conditions without background light



Tsuhao Chen

Conclusions

- Database is essential
 - ↳ Need to consider Lombard Effect
- Fusion is important
 - ↳ We can learn from acoustic ASR; we can perhaps lead ASR
- Confidence estimation is important
- Visual channel is not noise-free

Tsuhao Chen

Related Forums

- IEEE Multimedia Signal Processing (MMSP) Technical Committee, 1996~
- *Proceedings of IEEE*, Special Issue on MMSP, 1998
- IEEE MMSP Workshops
 - ↳ Princeton 1997, Los Angeles 1998, Copenhagen 1999, Cannes 2001, St. Thomas 2002
- IEEE International Conf. on Multimedia and Expo. (ICME)
 - ↳ New York 2000, Tokyo 2001, Lausanne 2002, Baltimore 2003
- *IEEE Transactions on Multimedia*, March 1999~
 - ↳ Special issues: networked multimedia 2001, multimedia database 2002, multimodal interface 2003

Tsuhao Chen

Advanced Multimedia Processing Lab

Please visit us at:

<http://amp.ece.cmu.edu>

Or, please email me at
tsuhan@cmu.edu

Tsuhan Chen

Author Index

Aleksic, Peter,	
Billington, Scott,	9
Blimes, Jeff,	
Blue, Misty,	59
Chan, Michel,	27
Chen, Ken,	32
Chen, Steve,	1
Choi, Seunggho,	49
Chu, Stephen,	17, 32
Chuckpaiwong, Ittichote,	9
Fisher, Francis,	1
Fisher, Pete,	21
Garg, Ashutosh,	32
Geisheimer, Janathan,	9
Gowdy John	41
Greneker, Eugene,	9
Gurbuz, Sabri,	41
Hasegawa-Johnson, Mark,	32
Heinrich, Jason,	13
Huang, Thomas,	17, 32
Jing, Zhinian,	32
Kim, Gwang-Myung,	53
Kim, Jinyoung,	49
Kim, Jung H.,	13, 53
Levinson, Stephen,	32
Li, Danfeng,	32
Lin, DongCheng,	53
Lin, John,	32
Nakamura, Satoshi,	37
Neti, C.,	62
Ntuen, Celestine,	59
Omar, Mohamed,	32
Park, Seongmo,	49
Potamianos, G.,	62
Scanlon, Mike,	1
Taylor, Justin,	13
Tsuhau, Chen,	76
Waibel, Alex	67
Wen, Zhen,	32
Yoon, Sung H.	13, 53